Irina Kuznetsova

Dual codon usage in germline variants of prostate cancer genomes

Master's thesis

Tampere, Finland, March 2015

Supervisors: Professor Matti Nykter
M.Sc. Tommi Rantapero

University of Tampere
Institute of Biomedical Technology
Department of Bioinformatics

UNIVERSITY
OF TAMPERE

Acknowledgement

During graduation ceremony from my first degree Professor S.S. Mnatzakanov gave us a valuable advice “never stop learning”, which I am following since that day. I could not imagine that the study of bioinformatics will change my life so dramatically. Moreover, I found my passion and I like the challenge provided by the field. Frankly, it was the most difficult study time in my life, but through challenges I achieved the results and I am incite to go only ahead.

The professional vision of science, well planned study process and excellent research team of Mattin Nykter gave me a very solid knowledge of the Bioinformatics field. I would like to express my gratefulness to Professor Matti Nykter and PhD student Tommi Rantapero for they constant support with the thesis, for providing a flexibility and understanding during unexpected big changes in my life. Further, I would like to say many thanks to each member of the research group, as each person delivered a friendly and relaxing environment and never refuse in providing an explanation or help. Finally, this is a very good chance to express the gratefulness to my family for they non-stop support and love.

I am thankful for being surrounded by a lot of great people who makes a positive environment at my work and my private live. Thus, I am capable fully contribute myself to the loved job.

Master's Thesis

Place:	University of Tampere
	Master's Degree Program in Bioinformatics
	Institute of Biomedical Technology
	Tampere, Finland
Author:	Irina Kuznetsova
Title:	Dual codon usage in germline variants of prostate cancer genomes
Pages:	77
Supervisors:	Professor Matti Nykter
	M.Sc. Tommi Rantapero
Reviewers:	Professor Matti Nykter
	Dr. Juha Kesseli
Time:	May 2015

Abstract

This thesis work builds on the recent discovery by Stergachis and his colleagues, which describes the process the genome uses to write genetic code. This work extends the previous vision in genome research, which states, that codons and regulatory elements work independently. Codons are a triplet of nucleotides that encode amino acids; and regulatory elements are responsible for regulation of the gene expression. However, as discovered by Stergachis and his colleagues in 2013, around 15% of codons within 85% of human genes are occupied by *transcription factor binding sites* (TFBSs) (see Stergachis et al., 2013). Consequently, these type of codons encode two types of information. They were labelled 'duons' and described as highly conserved entities with low levels of genetic variation. Overall, regulatory proteins bind to the same stretches of As, Cs, Ts and Gs and influence the process of gene expression, and also specify the amino acids of the protein that is made. This work applies Stergachis findings of 'duons' to analyse a variant data.

An interesting fact of Stregachis work is that a mutation may occur without affecting a protein. This happens due to the ability of some amino acids to be encoded by a multiple

combination of nucleotides (codons). Obviously, if an alteration occurs in the codon, which still encodes for the same amino acid, the functionality of the produced protein remains the same. In this case, *transcription factors* (TFs) bind to an altered (mutated) region, implicating a change of activity of TFs due to the fact, that the genetic pattern has been modified. As a result, wrong instructions are given to the expression of a gene, as Stergachis and his colleagues discovered (Stergachis et al., 2013). His discoveries led to the finding, that 13% of the *deoxyribonucleic acid (*DNA) mutations leading to a disease development are located in 'duons'. Thus it is important to investigate disease-associated variants within 'duons' that increase the risk of disrupting both regulatory and protein-structural function.

A finding by Kircher in 2014 - the application of a method that aimed at the interpretation of pathogenicity of human genetic variations – lead to a new method. This method developed by Kircher in 2014 is called the *combined annotation dependent depletion* (CADD) tool. It uses a single C score to annotate a variant as pathogenic. In contrast to other methods the CADD takes into consideration regulatory elements, thus the CADD tool was selected for this project work.

These two research findings are used in the thesis work. The goal of this work was therefore the extraction and recording of variants from provided data, which have potential for 'duons'. To achieve this goal, the thesis applied the techniques of the C score, the *position weight matrix* (PWMs), and p value estimation. The aim of this study was to apply the PWMs framework, and C score on provided data, in order to extract and record those variants from the data that have potential for' duons'. Thus they could be putative causes of a disease development. First of all, the provided data was filtered to identify pathogenic variants based on C score. Afterwards, the above presented concept was used to compute the TFBSs for original reference and mutated nucleotide sequences, where the maximum and minimum difference between these scores were found and used as a criteria for computing p value. Eventually, the resulting set of genes was submitted to the *Kyoto Encyclopedia of Genes and Genomes* (KEGG) pathway database and analysed for correlation of mutations to the type of a disease.

The outcome of the KEGG database analysis represents the main pathways where resulting genes are involved into metabolic, cancer, and neuroactive ligand-receptor interaction pathways.

Contents

ABBREVIATIONS

A	Adenine
ATP	Adenosine Triphosphate
bp	Base Pair
CADD	Combined Annotation-Dependent Depletion
cAMP	Cyclic Adenosine Monophosphate
ChIP-seq	Chromatin Immunoprecipitation-Sequencing
DNA	Deoxyribonucleic Acid
DNase	Deoxyribonuclease
dsDNA	Double-Stranded Deoxyribonucleic Acid
ENCODE	Encyclopedia of DNA Elements
GERP	Genomic Evolutionary Rate Profiling
GWA or GWAS	Genome-Wide Association Study
HGP	Human Genome Project
KEGG	Kyoto Encyclopedia of Genes and Genomes
LCR	Locus Control Regions
LRT	Likelihood Ratio Test
MIP	Molecular Inversion Probe
mRNA	Messenger Ribonucleic Acid
MT	Mutation Tester
NNs	Neural Networks
No.	Number
nsSNV	Nonsynonymous Single-Nucleotide Variations
PCR	Polymerase Chain Reaction
PFM	Position Frequency Matrix
phyloP	Phylogenetic P-Values

Poly(A)	Polyadenylation
PolyPhen	Polymorphism Phenotyping
PWMs	Position Weight Matrices
RE	Response Elements
RFs	Random Forests
RNA	Ribonucleic Acid
SIFT	Sorting Intolerant From Tolerant
SilVA	Silent Variant Analyser
SNPs	Single Nucleotide Polymorphisms
SNV	Single Nucleotide Variant
SVM	Support Vector Machine
T	Thymine
TF	Transcription Factors
TFBSs	Transcription Factor Binding Sites
TFIIIA	Transcription Factor IIIA
U	Uracil
UCSC	University of California at Santa Cruz (Genome Browser)
UniPROBE	Universal PBM Resource for Oligonucleotide-Binding Evaluation
UV	Ultraviolet
VEP	Variant Effect Predictor

LIST OF FIGURES

LIST OF TABLES

1 Introduction

One of the main challenges in biomedical research is to establish the association between genomic variation and phenotypic differences. It is believed that genetic variations play the main role in human diversity. Gene expression is the process which uses genetic information for proteins synthesis. Any dysregulation in this process will affect cells' responses to the environment, cells communication and eventually will lead to development of diseases (Bryois et al., 2014).

According to the Stergachis and colleagues the genome contains two codes (genetic code and regulatory code) that function collectively. The sequence of amino acids in a protein is represented by the genetic code, while regulatory code is responsible for specification of the recognition site for TFs (Stergachis et al., 2013). Thus, alteration in the sequence might lead to the production of the same protein, but alteration of function of binding TFs, that eventually will incorrectly instruct the gene expression process. Moreover, it was observed that 13% of human exons contain 'duons', so in the thesis work it was of high interest to investigate the exons regions of the genome (Stergachis et al., 2013).

High-throughput sequencing techniques are used for producing a vast amount of genetic variants data. Exome sequencing method was represented as a technique capable of extracting only protein-coding regions out of the whole genome. Thus, it is considered as a time and money efficient method.

As the term 'duons' is related to a dual work of human codons, the goal of the thesis work was to estimate how binding of TFs to mutated region of a sequence alter TFs instructions, which genes could be affected by this event in provided data. The expected outcome was a set of putative genes.

Overall, the understanding of occasions that cause a disease is exacerbated by complexity of events involved into this process. Thus it was important to examine and compute the TFs binding score and output the set of genes that could be provided by wrong instructions and produce a wrong protein.

2 Literature Review

2.1 The Genetic Material of an Organism

The *Human Genome Project* (HGP) was an international research effort; that had the aim to determine sequences of the human genome that construct this genome. The achieved results were stunning as for the first time information about human genes structure, its organization and functions was accomplished. This type of information is essential for understanding human being, and has a major impact in medicine and biotechnology fields.

Genome performs a complex hereditary material that consists of an organism's genetic instructions encoded as DNA sequences strung together in 23 chromosomes pairs. This information is housed in a complex form of nuclear genome, which is 99.9995% of genetic information and a simple mitochondrial genome that is remaining 0.0005% (Strachan T, 1999).

Figure 2.1 illustrates two main keepers of human genetic information, which maintain different amount of genetic information. Majority of genetic information is stored at the nuclear genome. These two genomes are different in its structure.

The mitochondrial genome is simpler in its architecture than nuclear genome. It is represented as one chromosome in a form of circular double-stranded DNA, it codes only for specific proteins that are generally used for mitochondria metabolic processes, such as an *adenosine triphosphate* (ATP) synthesis, fatty acids metabolism. The important fact about mitochondrial genome is that it is maternally inherited. Consequently, independently from the gender of an offspring the mitochondria genome comes only from a mother. The human mitochondrial genome contains *double-stranded deoxyribonucleic acid* (dsDNA) molecule, which encodes for (Figure 2.1) 13 polypeptides of oxidative phosphorylation system and 22 and 2 mitochondrial *ribosomal ribonucleic acid* (rRNA) that belongs to *ribonucleic acid* (RNA) machinery (Strachan T, 1999).

Table 2.1 Differences in human nuclear and mitochondrial genome (Strachan T, 1999)

	Nuclear genome	Mitochondrial genome
Size	3300 Mb	16.6 kb
No. of different DNA molecules	23 (in XX) or 24 (in XY) cells, all linear	One circular DNA molecule
Total no. of DNA molecules per cell	23 in haploid cells 46 in diploid cells	Several thousand
Associated protein	Several classes of histone and nonhistone protein	Largely free of protein
Number of genes	~65 000–80 000	37
Gene density	~1/40 kb	1/0.45 kb
Transcription	The great bulk of genes are transcribed individually	Continuous transcription of multiple genes
Introns	Found in most genes	Absent
% of coding DNA	~3%	~93%
Recombination	At least once for each pair of homologs at meiosis	Not evident
Inheritance	Mendelian for sequences on X and autosomes; paternal for sequences on Y	Exclusively maternal

At the same time the genome may be organized within its nuclear environment, defined as a linear double stranded cellular DNA. The number of genes holds in nuclear genome remains unknown. However, some studies estimated around 80,000 genes (Strachan T, 1999). In contrast to mitochondrial genome which mostly contains (approximately 93%) of the DNA sequence protein-coding regions, nuclear genome has only around 2% of such regions. Table 2.1 illustrates other differences of mitochondrial and nuclear genomes (Strachan T, 1999).

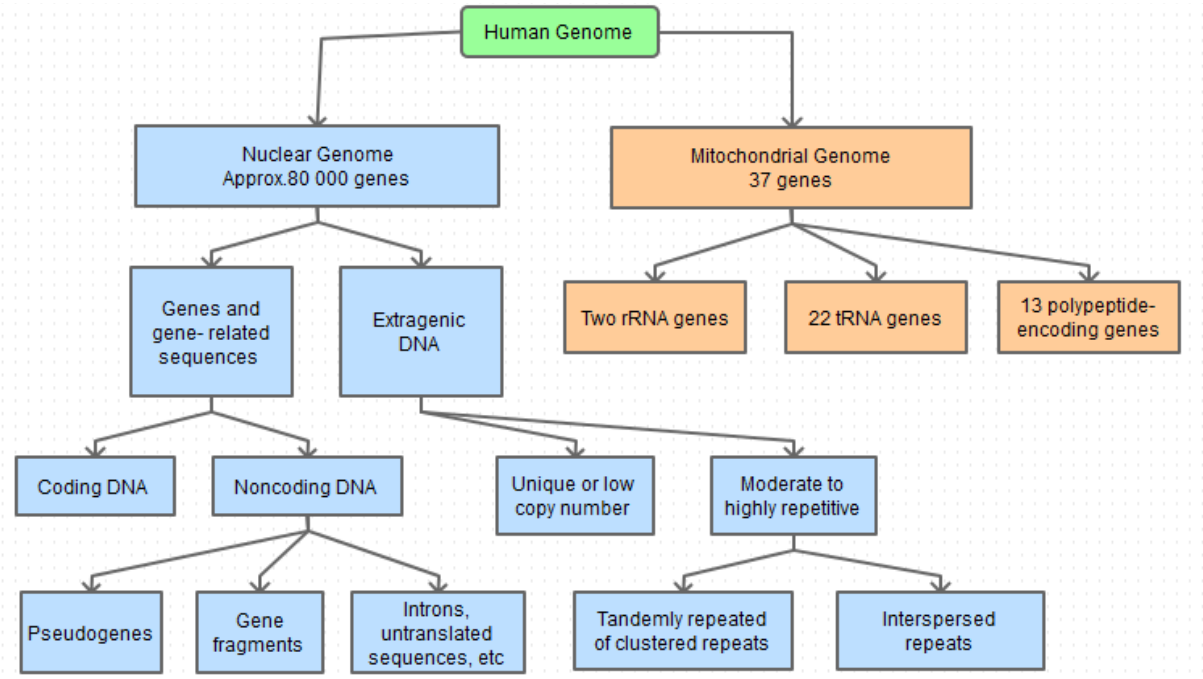

Figure 2.1 Human genome organization (Strachan T, 1999)

Inside of each cell there is a nucleus that headed all genetic information in a form of chromosomes, and regulates activities of cells (Figure 2.2). Chromosomes are made up of genes. The process, called DNA packaging allows an enormously long DNA molecule easily to fit into a chromosome. During this process DNA is tightly looped, coiled and wrapped around proteins, called histones. All genetic information is encoded in a long double helix shaped DNA molecule, built with four chemical building blocks: adenine (A), guanine (G), cytosine (C), and thymine (T). A DNA sequence is a random order of nucleotides, which are organised in triples, that eventually assembly into various amounts of complexes, called genes. There are approximately 20,500 protein-coding genes in the human genome (Bolsover, Shephard, White, & Hyams, 2011). The central dogma of molecular biology is interconnected with complex series of events starting from production of RNA from DNA and turning on to a final product protein that regulates cells functions (“An Overview of the Human Genome Project,” n.d.).

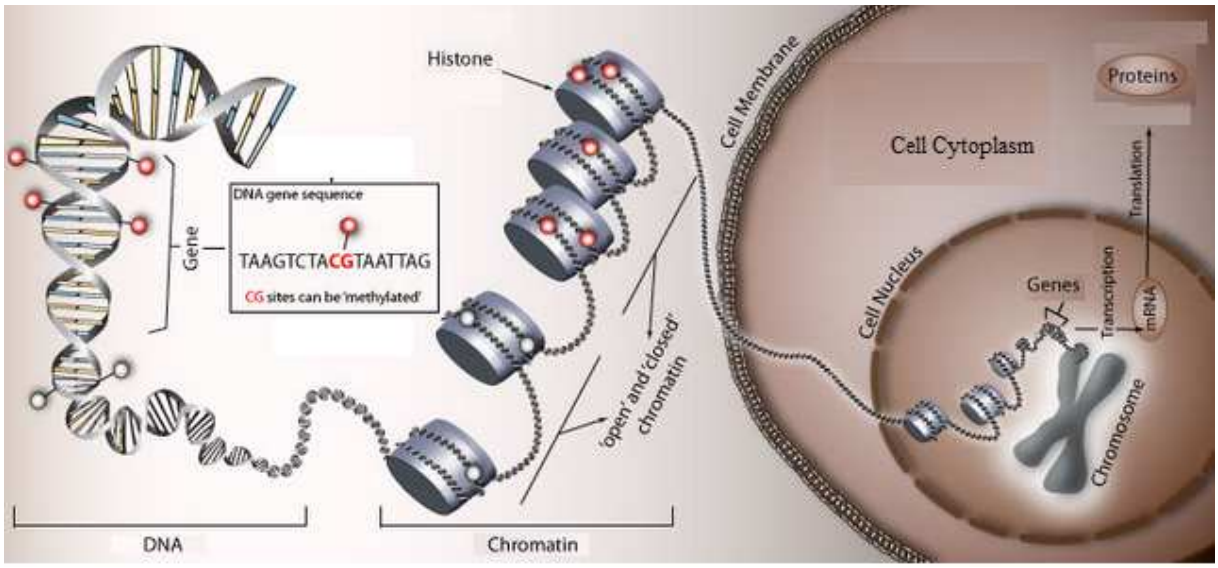

Figure 2.2 The genome inside the cell ("AEpiA :: Australian Epigenetic Alliance," n.d.)

The DNA molecule can be divided into several regions such as protein-coding (around 2% of human genome), that codes for proteins and non-coding (around 98%). Although the amount of noncoding regions is huge (approximately 99%), noncoding DNA sequences have important biological functions as well, such as transcriptional and translational regulation of protein-coding sequences. On the other hand remaining 1% of DNA protein-coding regions (exons) is essential for understanding a disease causing reasons. Development of massively parallel sequencing technologies enable of sequencing regions of the main interest. The exome capture technique permits to extract only protein-coding region. In addition although the mitochondrial genome has smaller and simpler structure than nuclear, it also represents one of the vital roles of genetic system. Thus, mutations at mitochondrial genome play essential role in research and considered as the main cause of genetic disorders (Taylor & Turnbull, 2005).

2.2 Gene Expression

The human body contains about 100 trillion cells, which can be separated in several groups according to their functions. Each cell is responsible for fulfilling its duties at specific time, in certain quantity. A set of proteins that synthesized from specific genes provide a specific cell type with instructions. So that each group of cells know exactly what, when and in what

quantities it has to produce. Thus, gene expression is one of the complex and important processes in the human body that uses information encoded in a gene for production of a protein. Accordingly the fundamental dogma of molecular biology is that proteins are produced from DNA through RNA. More precisely the DNA is transcribed into the *messenger ribonucleic acid* (mRNA) inside the nucleus, and afterwards mRNA migrates into the cytoplasm and each mRNA molecule is translated into proteins. Obviously any dysregulation during this process may lead to formation of a disease (Alberts et al., 2002). Gene expression is a sequential complex flow of different processes. Major steps of this process are illustrated on the Figure 2.3.

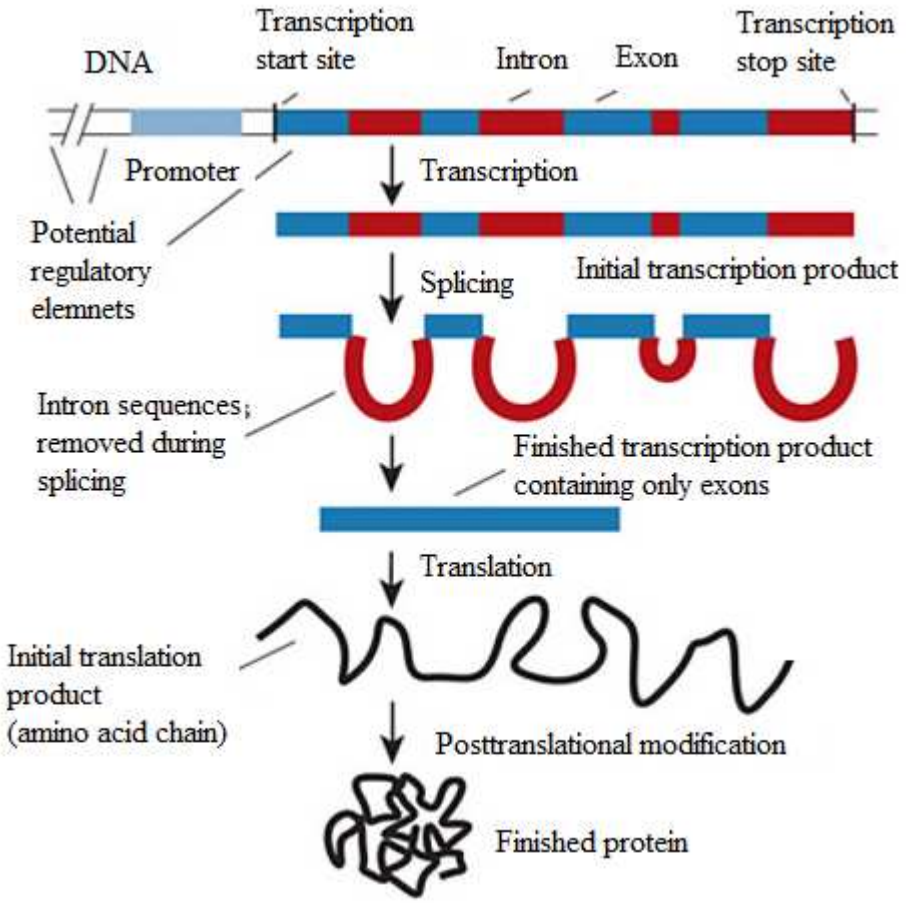

Figure 2.3 Gene expression workflow (Mandal A, 2015)

Complex flow of gene expression events starts from transcription. The transcription is the process when genes are copied and produce an RNA molecule, where noncoding regions (introns) are removed and mature transcript or mRNA is created. In order to initialize transcription process, enzyme RNA polymerase should be activated. Therefore TFs bind to

the core promoter (-30,-75,-90 *base pairs* (bp)) upstream from the transcription start site. After RNA polymerase is activated it binds to the promoter region of the DNA molecule (Figure 2.4).

Additionally, it is also known that RNA polymerase activity might be influenced by enhancer sequences that provide binding sites for regulatory proteins. Combination of regulatory elements and enhancer alter chromatin structure that consequently promotes or stops RNA polymerase and TF binding (Clancy, 2008). The major role of RNA polymerase is to separate double-stranded DNA molecule, by breaking hydrogen bonds, afterwards adding complementary nucleotides. This process is called elongation. The distinctive feature in producing mRNA molecule is that instead of *thymine* (T) it contains of *uracil* (U), which is complementary to *adenine* (A). Furthermore, the RNA molecule is single-stranded non helical molecule. The finalization of building complementary strand might be terminated in different ways. It might terminate process until a polymerase reaches termination sequence, on the other hand it can involve a termination factor which is special protein. Afterwards the process of removing noncoding nucleotide regions (introns) begins, and coding regions, exons are spliced together**.**

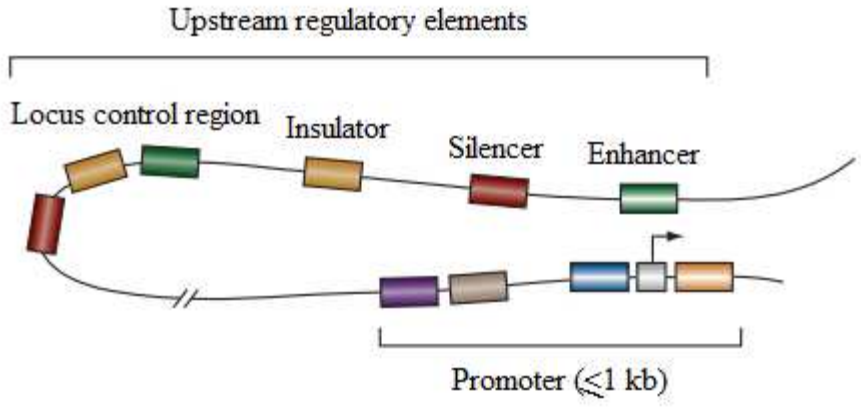

Figure 2.4 Gene regulatory elements (Maston, Evans, & Green, 2006)

The second major step during the gene expression is the process of manufacturing different proteins which is called translation, when the combination of three nucleotides, called codon is translated into 20-letter code of amino acids. The process begins in several ribosomal RNA molecules in complex with certain proteins that form ribosome. The initiation of this process starts when small subunit of ribosome binds to mRNA and searches for the start sequence

AUG (codes for methionine); afterwards a large subunit joins to form the complete initiation complex. The elongation process accumulates translated nucleotides until all of codons are read. The termination occurs when the complex reaches a stop codon (UAA, UAG, and UGA). Finally, produced protein is released.

Gene expression is a complex process that involves a lot of intermediate steps and interaction of biochemical elements such as genes, RNA molecules, and proteins (including TFs). There are varieties of different processes that cells are orchestrated for increasing or decreasing proteins production. It is clear that any disruption in this process might lead to the serious consequences, and eventually cause a disease.

There are three main steps that regulate the transcription stage: genetic, where control factors interact with genes; modulation, where control factors interacts with transcription machinery; epigenic, other factors than DNA alterations that affect transcriptions. In order to control the amount of mRNA translated into proteins, the post-transcriptional regulation adjust the capping, splicing, addition of the p*olyadenylation* (Poly(A)) Tail processes. The last major step in the gene expression is translation, this process mostly regulated at initiation stage.

In addition, gene expression process is highly error predisposed. According to D. Allan Drummond and Claus O. Wilke, alterations of nucleotides may be seen once in 1000 to 10000 translated codons. The other concept was proposed by same authors saying that the more intensively a gene is expressed, the higher chances to that a protein will be predisposed to the errors which are eventually affect organism's phenotype (Drummond & Wilke, 2009).

2.2.1 Structure of Protein-Coding Genes

Genome is represented as set of genes. Genes are made up of DNA, a long polymer sequence that is constructed from nucleotides. As the genome is continued entity there are parts that do not represent genes. These regions are called intergenic regions and they are genes separators. Thus a term gene is referring to any region of the genome that is essential for activation of biological functions of cells. There are several types of genes: non-transcribed regulatory genes transcribed RNA-genes and translated protein-coding genes.

Non-transcribed regulatory genes characterize sites for initiation and termination of DNA replication. Transcribed RNA genes produce RNA products like ribosomal RNA, transfer

RNA and etc. Finally, the last type called translated protein-coding genes codes for proteins respectively, which are illustrated on the Figure 2.4 (Laszlo P, 2009).

2.2.1.1 Promoters

A promoter is a DNA sequence that defines where transcription of a gene begins together with RNA polymerase. To initialize transcription RNA polymerase and TFs have to bind to a promoter region together. The other function of promoters is to define direction of transcription and indicate which DNA strand to transcribe. One prevalent type of promoters in eukaryotes is called TATA box, which is an AT rich sequence (consensus TATAA/TAA/T), located in 28–34 bp upstream of the transcription start site of a gene. Only about 24% of human promoters contain a TATA box, which is associated with tissue- or context-specific genes. The remaining 76% of promoters do not contain a TATA box and thus require another mechanism of initiation, which plays an essential role in connecting key elements during the transcriptional process (Sandelin et al., 2007).

There are other elements such as GC rich sequence (the Sp1 box) or the CCAAT box located in upstream of the promoter. The Sp1 box has ability to substitute TATA box main features in case of its absence and initialize transcription. All these elements are essential for starting transcription and failure one of them activates the action of another (Latchman, 2008).

2.2.1.2 Enhancers

Enhancers represent short regions of DNA sequence (50-1500 bp) which may be located at upstream, downstream, or within transcription regions. One of the distinctive characteristics of these elements is that they are located quite far from a transcription site. In addition, their major feature is ability to increase a rate of expression of a gene; on other words they reinforce the gene expression process. Enhancers might be tissue specific, whether they activate specific promoter of a specific cell. The other type of enhancers active in all tissues where it raise the level of gene expression in all cell types (Latchman, 2008).

2.2.1.3 Locus Control Regions

Genes that are presented on the same chromosome and located at a very close to each other position, as well as co-regulated by a common cis-regulatory element are called linked genes. These cis-regulatory elements are called *Locus Control Regions* (LRC). One of essential properties of the LRC is strong enhancer activity. The process is tissue specific, that might influence the mechanism of transcription machinery (Q. Li, Peterson, Fang, &

Stamatoyannopoulos, 2002). Lack of these elements contribute to the disruption of the normal way of gene expression or its full cancelation (Latchman, 2008).

2.2.1.4 Silencers

Silencers are elements that have opposite properties to enhancers and LCR, it is main task is to inhibit expression of certain genes. A silencer was discovered at specific genes, thus it is gene specific element. The activity of silencers is dual it may be fully active or tissue specific (Latchman, 2008).

2.2.1.5 Insulators

Insulators are elements that block interactions between enhancers and promoters that enable to act on the large distances. There are two mechanisms how insulators obstruct connections between these elements. They either affect on the chromatin structure, or prevent DNA from looping (Latchman, 2008).

2.2.2 Transcription Factors

2.2.2.1 General Information

TFs are proteins that behave in a similar way to a ‘membrane’ that permits a certain amount of genetic information to pass from DNA to RNA. The quantity of TFs present in the genome depends on its size. The larger the size of the genome is more TFs are present there. Furthermore, TFs are capable to work in cooperation with other protein complexes or without them. In addition, the gene expression is generally regulated by a combination of TFs which are typical arrangements for this process. TFs are complex biological entity that are involved in complex vital processes, like cells division and differentiation, metabolic and physiological balance and others (Latchman, 1997).

The gene expression represents a complex flow of various processes that generally are divided into two main steps: transcription and translation. The transcription is initialized by presence of TFs, when they bind to a DNA sequence. Therefore the biological function of DNA depends on a site where DNA binding proteins find targets.

Although the mechanism of TFs preferences in specific binding sites of the DNA sequence is not fully understood, there are various concepts that try to describe the specific selectiveness of TFs in binding. One of the concepts describes interactions between protein (TF) and the DNA sequence from structural point of view and can be divided into two subclasses: ‘base

readout' based on the recognition of specific chemical signs by the protein (Figure 2.5 -1); and 'shape readout' based on recognition of sequence-dependent DNA shape by the protein (Figure 2.5-2). These two mechanisms are considered as incentive forces that permits the TF to find a target (Rohs et al., 2010).

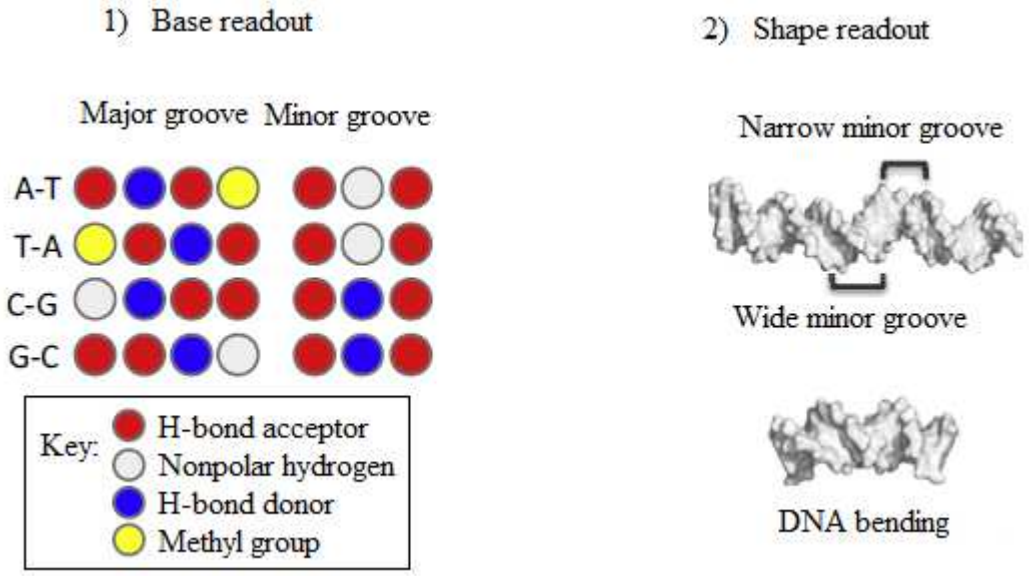

Figure 2.5 Types of structural interactions between TF and DNA (Slattery et al., 2014)

1) illustrates 'base readout' structural interactions between TF and DNA sequence in major and minor groove. Where the major groove has random distribution of the key elements, than in the minor groove the structural organization of the key elements is seen (Slattery et al., 2014)**.**

2) illustrates 'shape readout' structural interactions between TF and DNA sequence. The DNA sequence mostly has distorted shape that affects the electrostatic potential. (Slattery et al., 2014)

The other concept is built on computational methods which are aimed to model DNA motifs based on small experiments (like *deoxyribonuclease* (DNase) I footprinting) or simulated data. Microarray development enhances the amount of methods aimed to explain the way TFs find their targets.

According to Yongping Pan and his group *response elements* (RE) are essential player in TF binding mechanism. The strong affinity between TF and RE directs to the TF binding process (Pan, Tsai, Ma, & Nussinov, 2009). However, his group has found that concentration of protein factors is one of the essential conditions for further selection of the binding positions. The increase in concentration will lead to alter of allosteric properties and eventually to the protein structure. Consequently the protein will bind to the sequence position that is consistent with the DNA sequence. Same strategy applies for DNA sequence (Pan et al., 2009). However, the other factors that might influence the TFs binding affinity to a specific location could be poor connection with DNA backbone.

Once TF is bound to DNA it can activate or repress enzyme that controls translation, by turning on or off genes respectfully. A human body consist of various types of cells, which are regulated by different genes at different time. While genes that regulate liver cells are turned on, genes that regulate skin cells may be turned off. Similar scheme applies to a cancer affected region, where genes that have to be expressed are suppressed. Despite all cells contain the same genome they act differently, depending on cells type they represent ("Transcription Factor | Broad Institute of MIT and Harvard," n.d.).

2.2.2.2 TF families

The expression of various genes activates functions of different cells; likewise the expression of different types of genes is regulated by various TFs. There are four common groups of DNA motifs that can be allocated: zinc finger, the helix-turn-helix, the leucine zipper, and the helix-loop-helix motif.

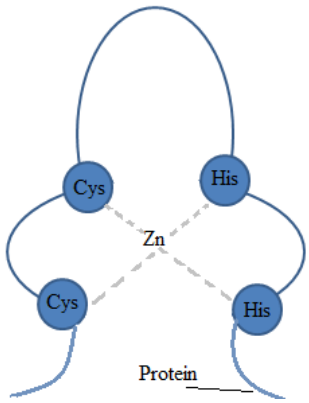

Figure 2.6 Zinc finger proteins

The zinc finger was first found 15 years ago in Xenopus *transcription factor IIIA* (TFIIIA) (Laity, Lee, & Wright, 2001). The name came from the zinc atom that builds the protein and uses it for tightly wrapping around. This motif represent a range of function including DNA recognition, RNA packaging, transcriptional activation, regulation of apoptosis, protein folding and assembly, and lipid binding (Laity et al., 2001). According to Alison Thomas, zinc plays a role in loop stabilization of this protein (Figure 2.6) by the R groups of two cysteine and two histidine residues (Thomas, 2013). Moreover, one side of the loops is represented as an alpha helix that is located in the major groove of the DNA sequence (Thomas, 2013).

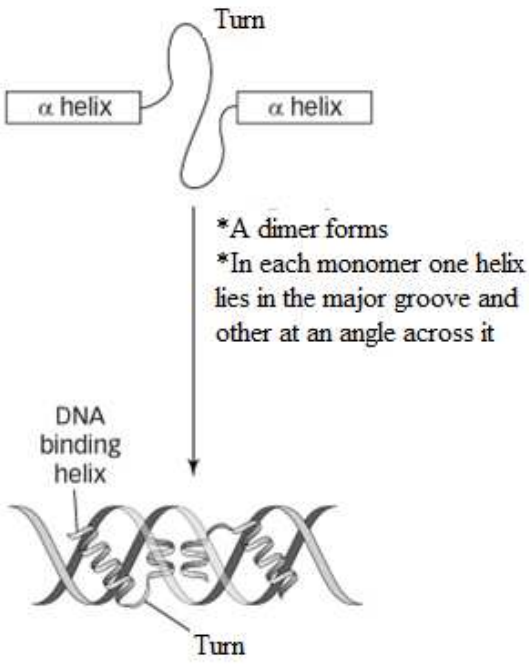

Figure 2.7 The helix-turn-helix (Thomas, 2013)

The helix-turn-helix (Figure 2.7) consists of two alpha helixes where both lies at an angle across DNA. Alison Thomas suggests that the amino acid R-groups of the C-terminal helix and bases operate with major groove and thus determine the selection of particular sequence location during binding (Thomas, 2013). Moreover, the helix-turn-helix contains homeodomain, paired box, forkhead and heat shock factors.

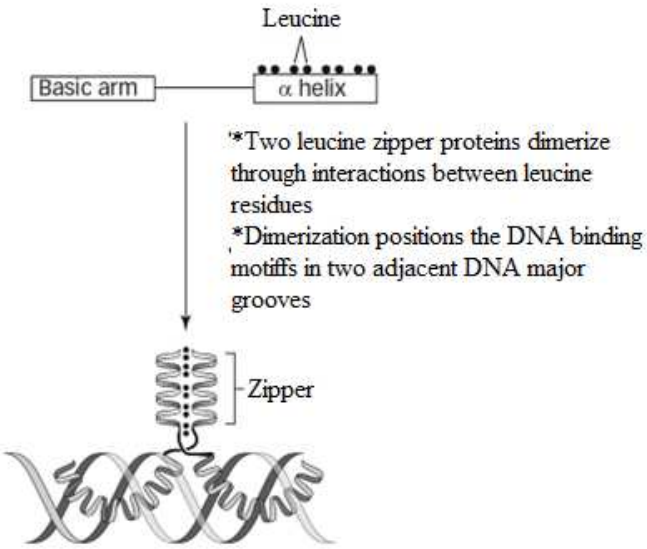

Figure 2.8 The leucine zipper motif (Thomas, 2013)

The leucine zipper motif (Figure 2.8) characterizes a family of TFs, that have alpha helical structure which is rich with leucine residues (every seven residues is leucine), that plays essential role in protein functioning. The protein is produced when two proteins 'zip' together and build a dimer, and the holding forces appear due to the connections of amino-acid leucine ("Atlas of Genetics and Cytogenetics in Oncology and Haematology," n.d.).

The helix-loop-helix motif, describes a family of TFs with alpha helices connected to a loop structure. It plays an important role in activating specific genes, and is similar in role to the leucine motif.

TF classes described above represent 80% of known TFs. TFs control the amount of transported genetic information from DNA to mRNA. They may also be characterized as positive or negative units by acting as activators or repressors respectively. Moreover, the process of gene expression is regulated more than only by TFs. The extracellular signals may turn on (or off) the gene expressions, as well as genes themselves have a power of regulating this process. One of the general features of binding proteins is frequent appearance of the same amino acids like asparagine, arginine, glutamine, glycine, lysine. Another is that binding occur through the major DNA groove. Finally, weak interactions such as hydrophobic, van der Waals forces, ionic bonds create strong bindings forces.

2.2.2.3 Mathematical Model of TF Binding

It is essential in the biomedical research to identify and describe the mechanism of binding TF to the DNA sequence, as this would provide the understanding of gene expression regulatory networks. The representation of TFBS model could be through biophysical view of this process. First of all, the process of binding to the DNA sequence is assumed to be reversible:

$$\mathrm{TF} + DNA \underset{K_{diss}}{\overset{K_{bind}}{\longleftrightarrow}} TF - DNA$$

Where,

S represents the rates;

K_{bind} (S) and K_{diss} (S) the sequence dependent rate constants;

E(S) is the binding energy (Djordjevic, Sengupta, & Shraiman, 2003).

$$\frac{K_{bind}(S)}{K_{diss}(S)} = K_{exp}(-\beta E(S))$$

Where,

$\beta = 1 / k_B T$

T absolute temperature

k_B is Boltzmann's constant

If we take into consideration the concentration (n_{tf}) of the provided solution with TFs, than probability of TF binding to sequence S is:

$$p(S) = \frac{K_{bind}(S)\, n_{tf}}{K_{bind}(S)\, n_{tf} + K_{diss}(S)} = \frac{K_{exp}\big(-\beta E(S)\big) n_{tf}}{K_{exp}\big(-\beta E(S)\big) n_{tf} + 1}$$

$$p(S) = f(E(S) - \mu) = \frac{1}{e^{(E(S)-\mu)/k_B T} + 1}$$

Where,

S, sequence;

n_{tf} concentration;

μ is the chemical potential (Djordjevic et al., 2003).

This formula illustrates the Fermi-Dirac distribution. This equilibrium describes that the TF will bind to a sequence only if the binding energy is below the chemical potential. At the same time if the binding energy is above the chemical potential the binding does not occur. It is assumed that the binding properties do not depend on the neighbouring nucleotides. So the binding energy could be written as:

$$E(S) \approx S\,\varepsilon \equiv \sum_{i=1}^{L}\sum_{a=1}^{4} \varepsilon_i^a S_i^a$$

Where,

ε_i^{α} shows the interaction energy of a nucleotide α at position i = 1... L;

$J_{ij}^{\alpha\beta}$ is the pair-dependent correction, is used as the parametrization for the sequence-specific interaction (Djordjevic et al., 2003).

2.2.3 ChIP-Seq

The ChIP-seq is the experimental way of finding TFBSs, than the PWM method is statistical approach. The major idea of both methods is identification of binding sites at the DNA sequence, but with utilization of different concepts. The understanding of how gene expression is regulated by proteins that bind to a DNA sequence plays an essential role in understanding many biological processes. The ChIP-seq is a powerful tool that is used for identifying the binding sites of TFs through the entire genome.

The chromatin is laid in the foundation of the method's name, which represents multifunctional molecule with properties of preventing the DNA from damage by fitting a long DNA sequence into a chromosome. It also controls the gene expression process and DNA replication, and tolerates mitosis after reinforcing the DNA. The general idea of the immunoprecipitation approach is ability to pull a protein by specific antibody, which is specifically attracted to this protein.

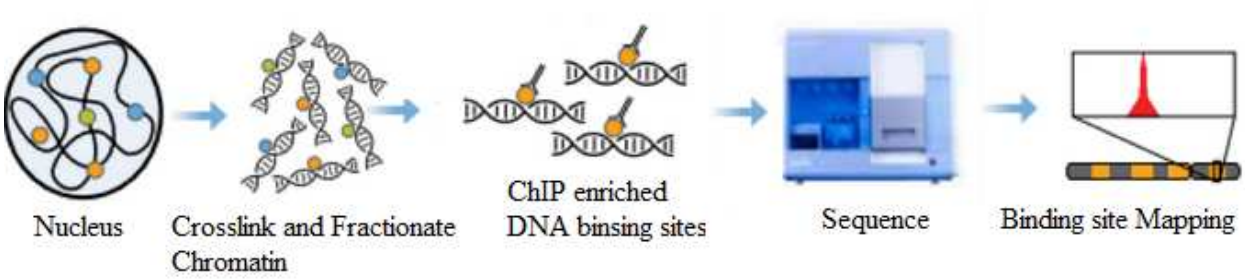

Figure 2.9 General ChIP-seq workflow ("Transcriptomics | Modeling Immunity," n.d.)

The ChIP-seq workflow is illustrated on the Figure 2.9. First of all, the ChIP-seq approach begins from the cross-linking process, which is capable of the histone modifications localization and also may define nucleosome position. The histone fragmentation begins after protein-DNA interactions are fixed. The length of the histone fragments should be in a range between 150 to 500 bp. An antibody that is specific to a protein of the interest begins the process of fragments enrichment of DNA-protein complexes. Finally, sequenced reads are aligned to the references genome with utilization of any alignment algorithm, such as BWA or Bowtie. Peaks can be analysed by using peak-calling algorithms, for instance MACS (Liu, Pott, & Huss, 2010).

The computational analysis of ChIP-seq takes into consideration the metrics of sequencing depth, quality checking, mapping, data normalization, assessment of reproducibility, peak calling, differential binding analysis, controlling the false discovery rate, peak annotation, visualization, and motif analysis (Bailey et al., 2013).

Consequently, the resulting data of ChIP-seq experiments varies from 100 to 10,000 predicted locations with resolution of around 50 bp (Wilbanks & Facciotti, 2010). The ChIP-seq technique has certain limitations:

- this method is still labour consuming
- the method allows to study one protein at the time
- it is limited by antibody specificity (Park, 2009)

However the output data depends on the quality of an antibody. A sensitive antibody makes detection of binding events easier. In contrast, there are certain advantages:

- this method offers higher base-pair resolution

- the hybridization step helps to avoid noise in resulting data
- the probe sequences limitations are not applied to the genome that makes analysis of the iterative regions easier (Park, 2009)

2.2.4 Methods Used for Predicting TFBS

2.2.4.1 General Information

There is an abundance of different methods, such as consensus sequence, PWM, position affinity matrix and k-mer that have been implemented for TFs binding sites (TFBSs) identification on DNA. The knowledge about TFBS provides better understanding of regulatory networks.

The main feature of TFs is to activate or repress the expression of genes by binding to specific sequence. Therefore, the ability to predict and identify TFBSs is the key point in understanding the gene regulation network. Moreover, it could help in understanding the influence of genetic variation on the process of gene expression disruption (Zhao, Ruan, Pandey, & Stormo, 2012).

The PWM is the quantitative approach used to predict TFBSs. PWM's are created based on finite number of experimentally derived motifs proven to be responsible for certain process like TF binding. The PWM for a DNA motif is represented as a matrix array with four rows named after nucleotides (A, C, T, and G) and the columns that represent the length of the binding sites. On the other hand the PWM for a protein motif may be performed as a matrix of 20 rows named after amino acids of a protein sequences (G, A, V, L, I, P, F, Y, W, S, T, C, M, N, Q, K, R, H, D, E).

The performance of the PWM approach considered as a quantitative model, for numerical representation of the binding sites at specific location on the DNA sequence (Mourad Elloumi, 2011). There are several methods that are used to construct PWMs. One of them is based on the experimentally determined binding sites (typically by the *chromatin immunoprecipitation-sequencing* (ChIP-seq approach) proposed by Staden (Nandi & Ioshikhes, 2012). The binding preference of TFs are not constant and vary from position to position. However, the same TF may express constant preference to the same position or variability. Consequently, observed binding sites are collected and stored at various databases, for instance JASPAR database (Mathelier et al., 2014).

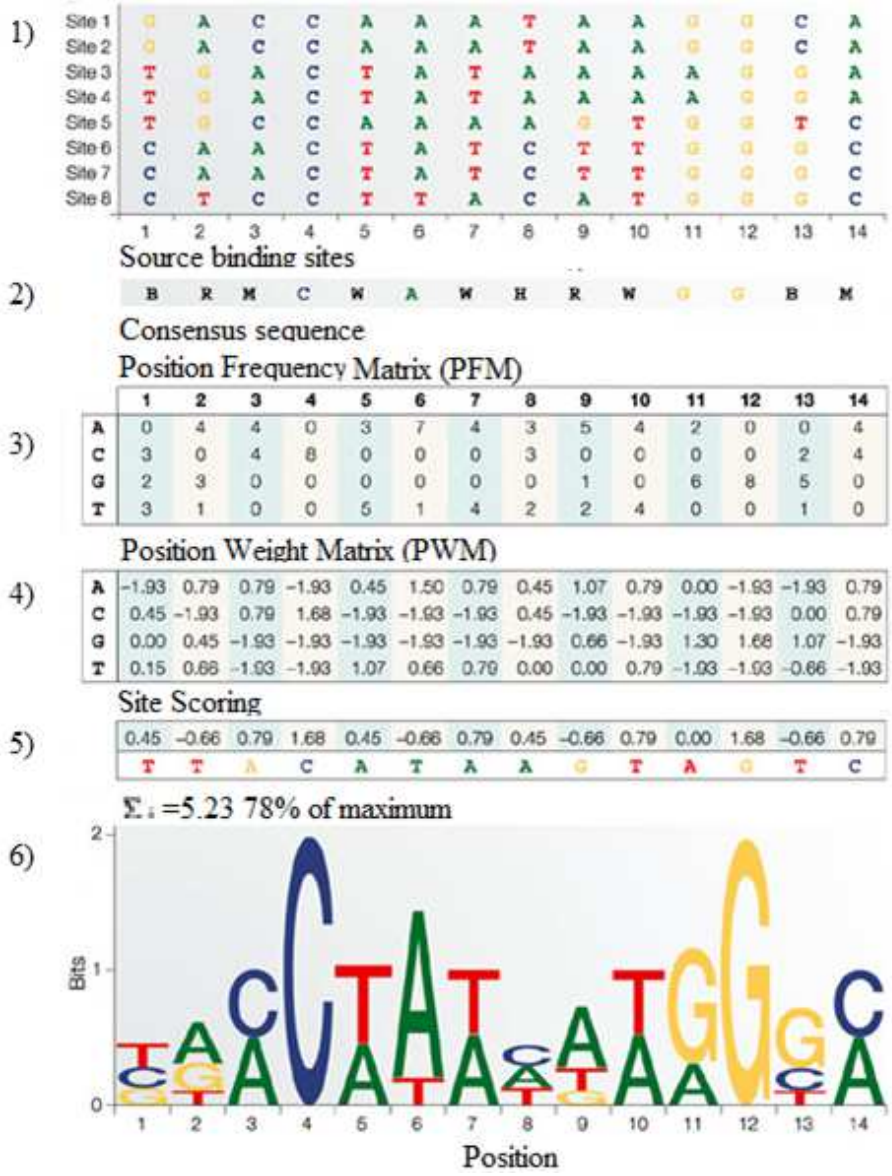

Figure 2.10 Building models for predicting TFBS (Wasserman & Sandelin, 2004)

One of the first steps for modelling TF binding sites is data collection. Such data could be simulated artificially in the laboratory conditions or derived through the utilization of high-throughput techniques, that allow to collect thousands of binding sites (Figure 2.10-1). Consensus sequences are one of the methods used for modelling TFBS. It shows the most frequent appearing residues among aligned sequences (Figure 2.10-2). Despite this method provides fast visual representation, it can't perform binding characteristics numerically (Wasserman & Sandelin, 2004).

2.2.4.2 PWM Construction

For constructing PWM (Figure 2.10-4), the first step is to align a large number of registered binding sites, and to calculate the relative frequencies of each nucleotide at this position the *Position Frequency Matrix* (PFM). The PFM shows the frequency of observed nucleotide at each position (Figure 2.10-3). If *p* is the PFM, when *p(b,i)* represents the number of counts of base *b* in position *i* of the alignment. The nucleotide probability is computed with equation (1).

$$(1) \quad \mathrm{p(b,i)} = \frac{f_{b,i} + s(b)}{N + \sum_{b' \in \{A,C,G,T\}} s\,(b')}$$

Where,

$f_{b,i}$ = counts of base b in position *i*;

N = number of sites; *p(b,i)* = corrected probability of base *b* in position *i*;

s(b) = pseudocount function (Wasserman & Sandelin, 2004).

The next step is to convert the PFM into a likelihood matrix. The elements of the PWM are calculated as log ratio of observed frequency divide by a relevant selected background model (equation 2).

$$(2) \quad \mathrm{W_{b,i}} = \log_2 \frac{p\,(b,i)}{p\,(b)}$$

Where,

p(b) = background probability of base *b*;

p(b,i) = corrected probability of base *b* in position *i*;

$W_{b,i}$ = PWM value of base *b* in position *i* (Wasserman & Sandelin, 2004) .

Each nucleotide of reference sequence matched to the PWM site is recorded and total sum is found (equation 3) (Figure 2.10-5).

$$(3) \quad S = \sum_{i=1}^{W} W\,l_i\,,i$$

Where,

l_i = the nucleotide in position *i* in an input sequence;

S = PWM score of a sequence;

w = width of the PWM (Wasserman & Sandelin, 2004).

In addition, the data can be performed visually called sequence logo (Figure 2.10-6). To compute the information content (in bits) in each position equation (4) can be used.

$$(4) \quad Di = 2 + \sum_{b} p_{b,i} \ \log_2 p_{b,i}$$

Where,
D_i = information content in position *i*;
p(b,i) = corrected probability of base *b* in position *i* (Wasserman & Sandelin, 2004).

Finally, the PWM method allows to get an accurate data, by taking into consideration mismatches by imposing position-specific penalties (Stormo, 2013). Moreover, the low level of both sensitivity and specificity is also provided by the PWM (Gershenzon, Stormo, & Ioshikhes, 2005).

2.3 Mutations

Frequently, term 'mutations' is associated with a process that has negative affects an organism features. However, mutations are common events that regularly occur in organisms and are linked to the human diversity.

Mutations can be distinguished from each other based on the modification it brings to a genome of an organism. The first type, considered as a harmful, and effect on the fitness of its host. The second typically have very small or no effect at all, called silent mutations. And the third type is advantageous, it leads to evolutionary advantage of certain phenotype (Keightley & Eyre-Walker, 2007). Nevertheless, mutations can also be described based on the place they occur. The event that leads to transmission of alterations to progeny is called germline mutations. It has been estimated that offspring receives around 100 new mutations from parents (Keightley & Eyre-Walker, 2007). On the other hand, mutations that affect only a host organism without being transmitted to an offspring are called somatic mutations (in non-reproductive cells). They are presented only in certain cells.

Another criterion that enable to group mutations is the length of affected nucleotide sequences. For short affected sequences the term gene-level mutations is used. Obviously it has impact on specific genes. In contrast the term chromosomal mutations is used to describe

mutations that alter longer regions of DNA sequence (“DNA Is Constantly Changing through the Process of Mutation,” n.d.). The human genome consists of coding and non-coding regions that both can be targets of mutations. However major interest for investigation is in the coding region, which can have two types of substitutions: synonymous and non-synonymous. While synonymous substitutions do not change the sequence of the gene product; the non-synonymous substitutions result on amino acids, with having various effect such as neutral, deleterious or positive (Strachan T, 1999).

Single-base substitutions or point mutations, exchange one nucleotide base to another. Clearly this type of mutations belongs to the gene-level and it includes three subclasses: missense mutations, nonsense mutations and silent mutations.

> **Missense mutations** (a type of non-synonymous substitutions), are types of mutation in which alteration of nucleotide in codon will affect the type of synthesized amino acid. This type of mutation has dual effect as some cases it has no effect at all, and then others might be deleterious. It is difficult to estimate the impact of this mutation on a disease development.
>
> **Nonsense mutations** (a type of non-synonymous substitutions), the alteration of nucleotide leads to a creation of stop codon (TAA, TAG, or TGA), that eventually terminates synthesis of a protein. Sequentially, the earlier the translation process stops the higher the chance to get non-functional protein.
>
> **Silent mutations** (a type of synonymous substitutions), the alteration of nucleotide in codon doesn’t change amino acid, as the same amino acid might be encoded by multiple combinations of nucleotides. The glycine, for instance is encoded by GGT, GGA, GGC, and GGG. Alterations at the third position lead to the production of the same amino acid, glycine.

Insertions and deletions mutations (a type of non-synonymous substitutions) are the other type of alterations that add or remove bp from the DNA of a gene, respectively are called frameshift mutations. The amount of inserted or deleted bp can vary from one to thousands. Frameshift mutations obviously lead to the different output of synthesized protein comparing to the possible output of original sequence (“DNA Is Constantly Changing through the Process of Mutation,” n.d.).

There are several factors that assumed to cause mutations in a DNA sequence. Some of them arise due to the effect of exogenous environmental factors such as *ultraviolet* (UV) radiation, chemicals, radiation and viruses. The other sources of mutations are endogenous; those are spontaneous errors during DNA replication and repair.

According, to Strachan, mitochondrial genome (Figure 2.1) is predisposed to mutations as well. There are various reasons for this event, one of them is that mitochondria genome has high amount of coding regions compared to the nucleus genome (Strachan T, 1999).

2.4 Exome Sequencing

Exome sequencing is a technique directed to a sequencing of all protein-coding regions (exome) of genome. The “EXpressed regiON” made the term ‘exon’, meaning there are regions that are translated or expressed as proteins (L. Eisenstadt, 2010). Exons represent a small part of the exome, so pieces of exons construct entire exome. Only 2% of the human genome are covered by gene coding regions, but significant amount (around 85%) of them are disease-causing (“Whole Exome Sequencing | Cost-effective analysis of protein coding regions,” n.d.). Studies prove that exome region represents highly enriched region of the genome, where variants have deleterious effect. Instead of sequencing a whole genome, as this process is time and finances consuming, the exome sequencing approach helps to identify only disease causing variants, found in coding regions of genes. The other benefit of this technique is unbiased examination (Teer & Mullikin, 2010).

2.4.1 Exome Sequencing Workflow

The exome capture techniques are used for isolating particular exome region from the whole human genome. The exome capture methods are based mostly on the idea of hybridization. The general workflow is represented on the Figure 2.11. First of all, the genomic DNA is fragmented, than the process of hybridization is applied. Fragments, that couldn’t undergo this reaction, are washed away. Captured regions of interest go through DNA sequencing and are analysed (Bamshad et al., 2011).

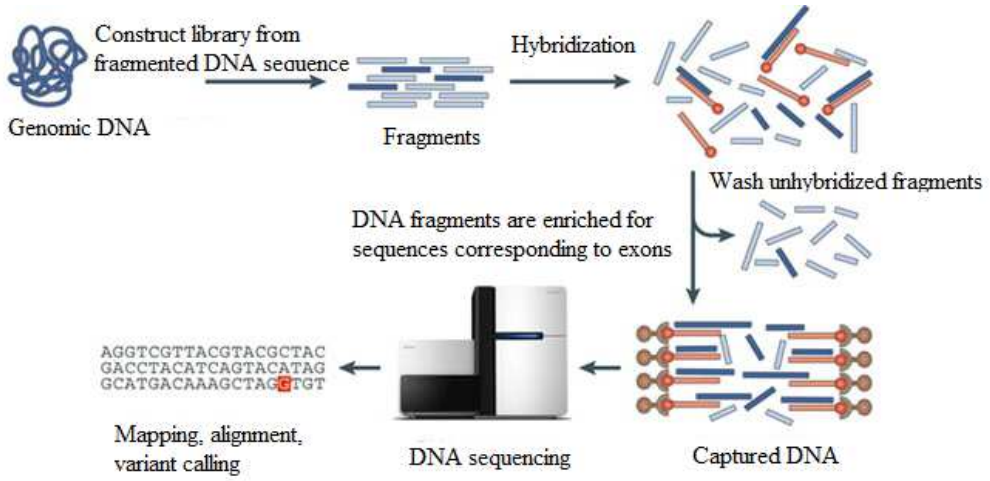

Figure 2.11 Exome sequencing workflow ("Box 1 : Exome sequencing as a tool for Mendelian disease gene discovery : Nature Reviews Genetics," n.d.)

Variety of methods exists for capturing genomic regions. They are characterized according to the used technique in capturing targets: *Polymerase Chain Reaction* (PCR), solid-phase capture and solution-phase capture methods, *Molecular Inversion Probe* (MIP) (Yoon et al., 2015).

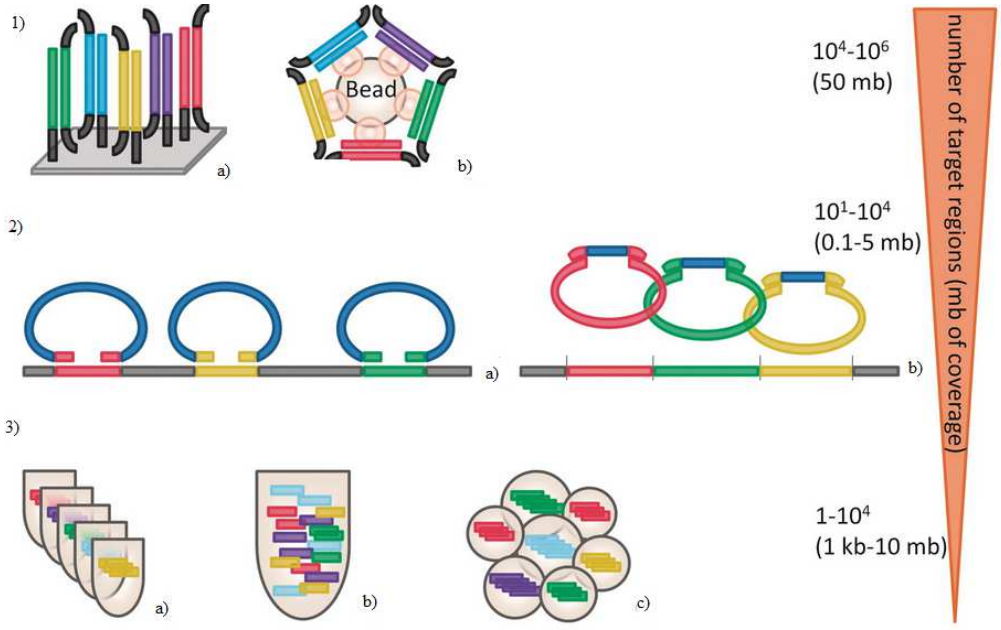

Figure 2.12 Commonly used target-enrichment methods (Mertes et al., 2011)

1) The hybridization target technique, where a) illustrates solid support, and b) in solution (Mertes et al., 2011)

2) Molecular inversion probes (MIP), where a) is a classical representation of MIP concept b) shows restriction enzyme cocktail (Mertes et al., 2011)

3) PCR enrichment where a) shows typical single-tube per fragment assay b) multiple PCR assay c) utilized for targeted enrichment (Mertes et al., 2011)

The hybridization technique (Figure 2.12 1)-a,b) is preferred for large target regions, can be performed by two various methods: reactions in solution (good for small amount of DNA) and reactions on a solid support (good for large target sets) (Mertes et al., 2011). The hybridization technique in solution is more efficient than on solid support. However, the main idea of these methods is based on hybridization of nucleic acid strands from sample data to the constructed DNA library, where fragments are complementary to the interested regions and capable of extracting an exome region (Mertes et al., 2011).

The second technique is called molecular inversion probes (MIP) or selective circularization (Figure 2.12 2)-a,b). MIPs are constructed with a gap, which is eventually hybridized by a region of interest and creates a circles structure (Mertes et al., 2011).

The last type of exome capture technique is enrichment by polymerase chain reaction (PCR). This approach use the main idea of DNA amplification (Figure 2.12 3)-a,b,c).

2.5 Mutations Effect

The main aim of the HGP was to provide a complete and accurate DNA sequence that build up a human genome. That was revolutionary approach as it opened a new way of utilization the DNA information towards large scale of investigations in biotechnology, disease causes, drug development ("An Overview of the Human Genome Project," n.d.). It is known now that the human genome of any individual is at 99% equal and only 1% makes humans different. That 1% is responsible for making living organisms look differently in shapes, sizes, weights, personalities, accessibilities to diseases and even abilities to tolerate a food (Norrgard K, 2008). Genes that construct a genome come as DNA sequences in a multiple versions. This means that the same gene might have slightly differences in DNA sequences between individuals. The eye colour is driven by the same gene, but with the slightly different variations in its architecture the variety of eyes colours is observed in different people. These variations are caused by mutations, so the final result is also different. Consequently, mutations are primary source to variations that occur randomly through the genome. The study of human genetic variation has both evolutionary significance and medical applications.

The HGP moved studies of a human genome forward. One of the leading methodologies is comparison of any human genome to the reference genome, produced by HGP. The easiest way to contrast genomes from different people is simply to compare them, as at some point possible inequality can be noticed. A *Genome-Wide Association Study* (GWA or GWAS) developed research technique that helps to identify genes that are involved in human diseases. The main idea of this approach is to compare genomes of several groups of people, healthy and carries of a studied disease, in order to identify regions of genome variations that possibly might lead to development of diseases. The examination of rare genetic variants would lead to the lack in associating to a disease. Thus the main focus of the GWAS is study of common genetic variants like *single nucleotide polymorphisms* (SNPs) or common *single nucleotide variants* (SNVs) that have been often associated to a disease ("Genome-Wide Association Studies Fact Sheet," n.d.).

The term SNPs defines as the single-nucleotide substitutions found throughout the genome that belong to members of one species that occurs in at least 1% of the population. It is important to understand that SNPs are not specifically localized in the genome, the appearance of SNPs might be found at any region of the human genome, for example in coding sequences genes, non-coding regions of genes, or in the intergenic regions (regions between genes). However, the utilization of GWAS approach helped to conclude that majority (around 88%) of SNPs were identified at intergenic or intronic regions (Edwards, Beesley, French, & Dunning, 2013).

The SNPs that occur in the coding region of the genome are divided on two types synonymous, that affect a protein, and non-synonymous that change a sequence of a protein. Non-synonymous are presented in two types as missense and nonsense. Since mutations are any changes in DNA, SNPs can be considered as mutation, which is presented at specific location of a genome in many peoples, and the presence of these alterations eventually leads to the human diversity. Therefore SNPs considered as evolutionary drivers, but not the cause of diseases. However, the combination of different SNPs in various genes may influence the risk of a single disease (DeWeerdt, 2004). Hopefully, the understanding of interactions and influences of variations could help for the further understanding how this events contributed to the predisposition to common diseases such as heart disease, diabetes, and various forms of cancer (Norrgard K, 2008).

The exome sequencing method is significant for investigating Mendelian diseases. It has been proved that the most common cause of Mendelian disease is the *non-synonymous single-nucleotide variant* (nsSNV) (M.-X. Li et al., 2013). Whether the term *single nucleotide variant* (SNV) is similar to SNP, which is common nucleotide alterations that are observed in population.

After all, the coding regions give a wide range of directions to explore causes of different diseases. The mutational landscape of tumours may be defined by focusing on somatic and germline SNVs.

2.5.1 The Main Concepts Used to Predict Variant Pathogenicity

With the help of the DNA sequencing technology scientists know precise arrangements of nucleotides in the genome. This knowledge helps in identifying disease-associated genes, which can normally be seen as the alteration in a sequence. However, such changes may be the representation of human diversity or causes of disease. Moreover, sometimes the rare variants can be presented in healthy humans, so the task of variant differentiation remains the main challenge in the bioscience (Ruklisa, Ware, Walsh, Balding, & Cook, 2015).

The process of assigning the right label to the discovered variant can be done through experimental analysis by applying a suitable system. Nevertheless, this is time, labour and money consuming technique. Consequently, an enormous amount of methods has been developed to recognize variants as harmful. These methods can be categorized in various ways. Some of them are based on supervised machine learning approach, while others on unsupervised machine learning; another type of tools are based on statistical approaches, while others use heuristic scores; some of methods use phylogenetic relationships and others pairwise comparison (Pollard, Hubisz, Rosenbloom, & Siepel, 2010).

The source of evidence that represent the pathogenicity can be allele frequency (definition: proportion of seen allele among all allele copies being considered (Cheung et al., 2000)), amino acid conservation (definition: a base sequence in a DNA molecule (or an amino acid sequence in a protein) that has remained essentially unchanged throughout evolution (“Glossary,” n.d.)), predictors based on physicochemical properties, and gene- and domain-specific effects (Ruklisa et al., 2015).

There are different ways how altered variant can affect a protein functions and lead to the risk of disorders. Some rules in predicting a variant of having harmful properties have been established experimentally:

- The location of a variant is characterized in SWISS_PROT database as binding site, active site, or involved in disulphide bond
- The variant has not suitable features to the family of homologous proteins
- Hydrophobic properties of a protein can be disrupted by a variant (Sunyaev et al., 2001)
- A variant can affect electrostatic properties
- A variant might affect dissolubility of a protein
- A variant might destroy protein ligand interactions (Sunyaev et al., 2001).

Furthermore, the online predicting programs utilize above mentioned features to predict variant deleterious properties. Mainly they can be divided into three groups: sequence and evolutionary conservation-based methods; protein sequence and structure-based methods; supervised learning methods.

Sequence and evolutionary conservation-based methods are based on amino acid conservation knowledge, used multiple sequence alignments and scoring functions. It's found that disease-associated variants are correlated to conservation concept. On the other hand, the output depends very much on the provided multiple sequence alignment. Tools that are constructed based on these concepts are for instance the *Sorting Intolerant From Tolerant* (SIFT), and Mutation Assessor. ("Missense Prediction Tool Catalogue | NGRL Manchester," n.d.). More information concerning a work of these tools can be seen in the Table 2.2.

Protein sequence and structure-based methods are built based on the structure of the protein. The output data might be interpreted in a wrong way without sufficient knowledge of protein structure features. The *Polymorphism Phenotyping* (PolyPhen-2) is a common tool that uses this concept (Adzhubei et al., 2010).

Finally, supervised-learning based methods are common way of variant pathogenicity interpretation. These methods can include *Neural Networks* (NNs), the *Support Vector Machines* (SVMs) and *Random Forests* (RFs) and naive Bayes classifiers. First of all a data that is used as reference have to be defined, so the algorithm has the pattern. Secondly, variant features are evaluated by using conservation or protein structure characteristics. Finally, the

algorithm ‘learn’ how to distinguish difference between variants. These types of learning require a wide range of known pathogenic variants for getting correct output. Mutation Taster and CADD are typical tools that utilize these concepts (“Missense Prediction Tool Catalogue | NGRL Manchester,” n.d.).

These methods are powerful have a lot of benefits. However, drawbacks are present as well. First of all, conservation metrics are not allele specific, than protein-based tools can’t be used for non-coding variants. Secondly, supervised-learning methods are trained on known pathogenic variants.

Table 2.2 Tools used for pathogenicity detection

Method Name	Brief Description
SIFT	The sorting intolerant from tolerant (sift) method based on sequence homology, computes the likelihood that an amino acid substitution will have a negative effect on protein function. SIFT is useful in research for study the influence of mutations on protein function (Sim et al., 2012).
Mutation Assessor	Mutation Assessor (ma), is the server which capable to predict the functional impact of amino-acid substitutions in proteins (definition: MutationAssessor.org). The method works by employing multiple sequence alignment, partitioning for identification of conserved positions; computing conservation scores, a specificity score and comparison of them for identification of the functional impact score (“MutationAssessor.org /// functional impact of protein mutations,” n.d.).
LRT	The *Likelihood Ratio Test* (lrt) uses goodness-of-fit statistical technique. It compares probabilities between conserved areas of a sequence and a neutral model (Chun & Fay, 2009).
PolyPhen	Polymorphism Phenotyping (PolyPhen-2), the method is used to detect deleteriousness of variants, by computing Naïve Bayes probability, as an output it estimates false positive or true positive rates. There are two types of Polymorphism Phenotyping: pp2_hdiv and pp2_hvar. The difference in these methods is in training data, and also pp2_hdiv is used for evaluating rare alleles, than pp2_hvar used for differentiation of harmful mutation from all human variation (Adzhubei et al., 2010).

Mutation Tester	The *Mutation Tester* (mt), a free, web-based application for rapid evaluation of the disease-causing potential of DNA sequence alterations. The method uses information from the different biomedical databases; the key player of disease potential detection of an alteration is Bayes classifier. The advantage of this method lays in the best performance from speed and accuracy ("MutationTaster - documentation," n.d.; Schwarz, Rödelsperger, Schuelke, & Seelow, 2010).
CADD	Combined Annotation Dependent Depletion (caddgt10) determines the genetic variation through performance of the C score as the measure of variant harmfulness (Kircher et al., 2014).

The study of pathogenic variants has to work in tandem with other approaches such as statistical association between a variant and a disorder, or ranking variants found from the genome based on its functional effect (Buske, Manickaraj, Mital, Ray, & Brudno, 2013; Pollard et al., 2010). In order to pick up a correct method to detect and label variants as benign or harmful following features should be taken into consideration: type of input data, methods that originates a data, and the training properties of selected methods.

There are many different studies that took place in investigating geography of pathogenicity events that lead to pathogenicity. Various tools and approaches were developed to identify deleterious variants. Pauline C. Ng and Steven Henikoff in research article "SIFT: predicting amino acid changes that affect protein function" describe the SIFT tool as the source of predicting if alteration in DNA sequence affect the protein functions or not (Ng & Henikoff, 2003). Research results of Jaaxin Wu and Rui Jiang suggested using multiple predicting algorithms to increase the accuracy in naming variants as harmful (Wu & Jiang, 2013). On the other hand there is opinion that synonymous SNVs play a role in developing a disease by affecting the way proteins are merged together, their expression and eventually function. The *Silent Variant Analyzer* (SilVA) tool was developed by Orion J. Buske and his colleagues, which is atomized application used to predict harmful synonymous variants within human genome. It was concluded that there are two most convincing types of features, splicing information and sequence conservation, that are used for detection of harmful synonymous (silent) mutations (Buske et al., 2013).

2.5.2 Combined Annotation Dependent Depletion (CADD)

The existence of vast amount of tools for pathogenicity prediction mostly is based on one metric. The CADD combines a lot of different metrics into one score.

2.5.2.1 Algorithm Implementation

The first step in training the algorithm is to construct a variant-by-annotation matrix. There were two types of data used: experimentally derived allele frequency information (from 1000 Genomes and Ensembl Compara) and simulated data based on empirical model of sequence with CpG dinucleotide-specific rates and mutation rates. The annotation metrics like conservation (from *Genomic Evolutionary Rate Profiling* (GERP), *Phylogenetic P-Values* (phyloP)), functional genomic data (from DNase) and TFs binding, exon-intron distance, expression levels in studied cell lines and protein-level scores (from SIFT, PolyPhen) were used to generate annotations information with utilization Ensembl, *Variant Effect Predictor* (VEP), *encyclopedia of DNA elements* (ENCODE) and *University of California at Santa Cruz* (UCSC) Genome Browser. The same type of model was used to train possible substitutions. The annotations were used in training a SVM with a linear kernel. Consequently, a rank system was used to assign values from 1 to 99 to trained variants (Kircher et al., 2014).

2.5.2.2 Pros and Cons

There are quite many benefits of CADD utilization. First of all, the CADD tool combines multiple annotations into single C-score. Secondly, C score relates to allelic frequency, it can be used for analysing coding or non-coding variants. Thirdly, C score capable to distinct a rare allele from set of disease-associated alleles. Fourthly, C score can be associated with somatic cancer mutations. Finally, CADD tool demonstrates strong prediction properties in pathogenicity, deleteriousness and molecular functionality and can be used for exome or genome studies (Kircher et al., 2014).

CADD method also has limitations. First of all, the accuracy of the tool can be limited be local mutation rate, background selection or biased gene conservation. Secondly, C scores may omit the differences in selective intensity. Finally, the ability to predict deleterious features in noncoding regions still require improvement (Kircher et al., 2014).

3 Research Goals

The vision of the thesis was to identify and catalogue variants that contribute to pathogenicity

The objectives are:

- ✓ To examine provided variants, to find location of harmful regions
- ✓ To identify more specific set of putative target genes
- ✓ To use CADD tool for scoring deleterious variants
- ✓ To compare precomputed CADD results with results achieved by using statistical analyses
- ✓ To apply PWMs framework for TFBSs prediction
- ✓ To identify which variants from the provided data have the potential for 'dual codon usage'
- ✓ To catalogue harmful variants

4 Tools

4.1 Python

Python is a freely available, an open source highly readable programming language ("Welcome to Python.org," n.d.). In this work the Python was mostly used for writing scripts for processing data. The Python version 2.7 was used for running all scripts. Additionally the NumPy package was installed, which is powerful tool, used in scientific computing along with Python, allowed to work easily with N-dimensional array objects ("NumPy — Numpy," n.d.). The NumPy was used in computational part of this thesis to convert PWMs into arrays.

4.2 Unix

Unix is a computer operating system, that mostly has been used in computational part as intermediate steps, like filtering, ordering, file observation ("The UNIX System, UNIX System," n.d.). Some of the Python scripts were running from the Unix platform as well.

4.3 R: Statistical Analysis Tool

R is an open source programming language for statistical computing and graphics (“The R Project for Statistical Computing,” n.d.). R platform has been mostly used for the initial part of the project. It was used for creating a file with necessary data, as well as for graphical visualization of genes frequency.

4.4 Combined Annotation-Dependent Depletion Tool

The CADD is powerful scoring tool that identifies the genetic variation through performance of the C score as the measure of variant harmfulness (Kircher et al., 2014). It is freely available open source tool.

The CADD training algorithm is based on the SVM learning method trained on potential pathogenic variants, used to distinguish benign mutations from deleterious. The C score represented all various characteristics of disease causing mutations into one single score. These characteristics include conservation metrics, functional genomic data, TFs binding, transcript information like distance to exon-intron boundaries or expression levels in commonly studied cell lines; and protein-level scores. Basically it combines into one C score a lot of various metrics that are used by many other tools related to detection of pathogenicity, such as GERP, DNase, SIFT and many others. This feature makes the CADD tool extremely strong and accurate in resulting data.

The thesis computational part has been built based on the main idea of using C score, which enable to estimate the pathogenicity potential of a variant, the rest of irrelevant data was sorted out. The threshold was set to 20, meaning these are predicted to be the 20% most deleterious variants.

4.5 BEDTools

BEDTools is a tool which allows to users easily to work with genomics analysis tasks. In the thesis the following BEDTools functions were used: intersect, slop and getfasta (“bedtools: a powerful toolset for genome arithmetic,” n.d.).

```
bedtools INTERSECT –a <file> –b <file1> –wo
```

The intersect function allows to check overlapped areas of provided genome with reference genome collected from the Genome Reference Consortium ("Genome Reference Consortium," n.d.). The '–wo' symbol is responsible for the output data, which retrieves only overlapped region of provided sequence.

bedtools SLOP –i <BED> –g <GENOME> –b 25

The slop function allows to increase a size of an object by a required number of bp. In case of this work it was extended from both sides by 25 bp.

bedtools GETFASTA –fi <FASTA> –bed <BED> –f <output FASTA>

The getfasta function converts data into FASTA format. The information holds in the output file was chromosomal position, start and end coordinates of sequence.

4.6 JASPAR and UNIPROBE Databases

JASPAR and *Universal PBM Resource for Oligonucleotide-Binding Evaluation* (UNIPROBE) databases were used as the repositories of information for the PWMs. The JASPAR CORE is an open access database, which provides information about TF for eukaryotes found experimentally (Mathelier et al., 2014). The UniPROBE database was used as the resource which contains information of PWM (Hume, Barrera, Gisselbrecht, & Bulyk, 2015).

4.7 KEGG

Kyoto Encyclopedia of Genes and Genomes (KEGG) is database resource which contains information about genomes, biological pathways, diseases, drugs and chemical substances. In the thesis work the KEGG database was used as pathway mapping tool in order to find possible associations of genes to disease ("KEGG: Kyoto Encyclopedia of Genes and Genomes," n.d.).

5 Methods and Approaches

5.1 Computational Analysis Overflow

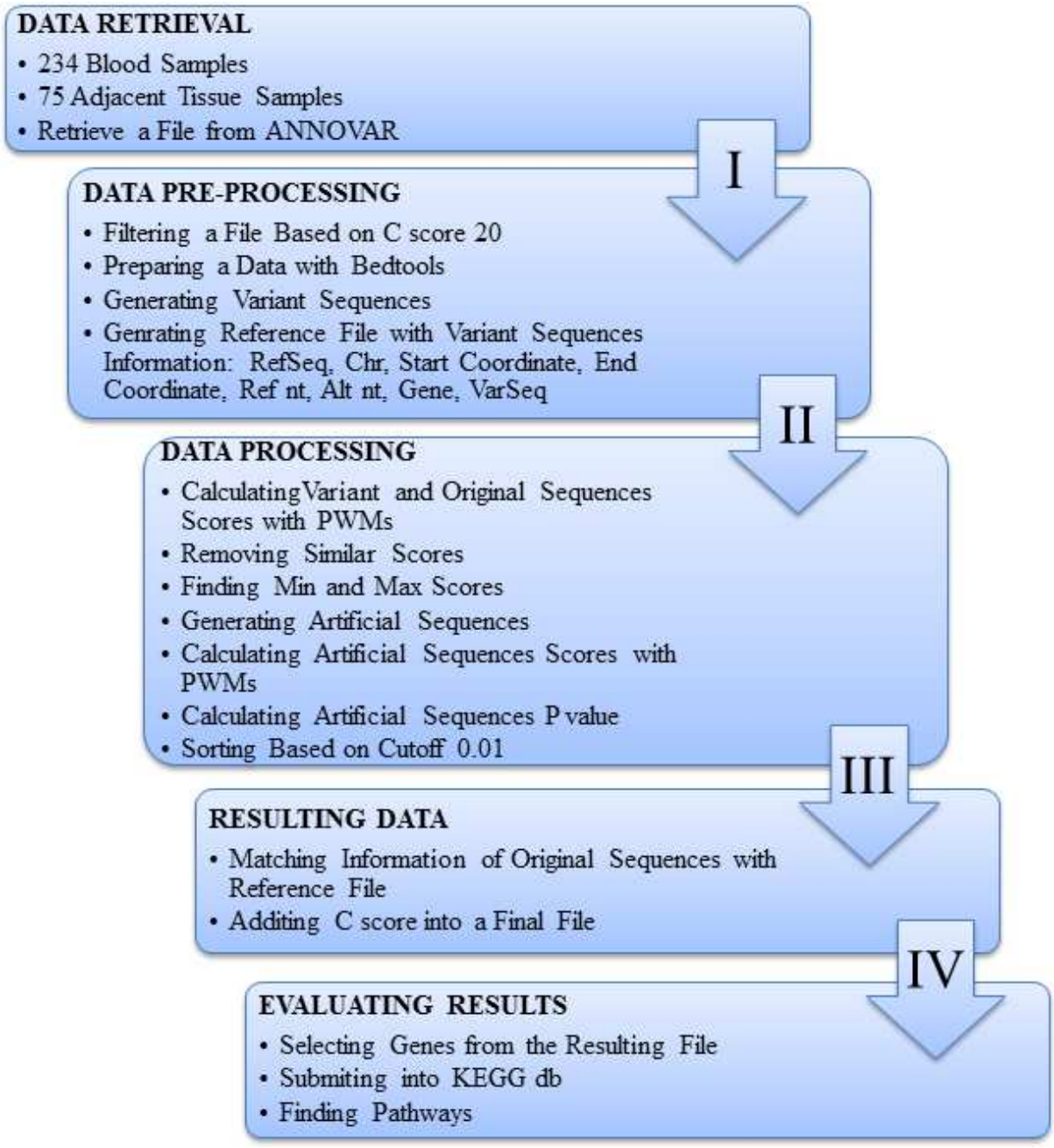

Figure 5.1 C computational analysis main steps

Figure 5.1 gives a full overview of the data processing during the computational part. Moreover, there were several main concepts, used for implementing this part: C score, PWMs score and p value concept. The final data was evaluated.

5.2 The PWM Concept

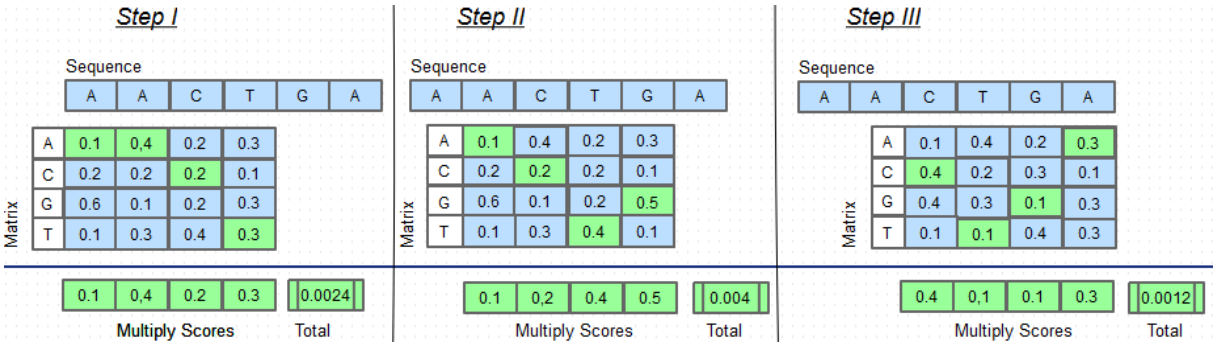

Figure 5.2 The main idea of PWMs utilization

Figure 5.2 describes the main idea, which was used in computational part for implementing the PWM concept. Based on this idea the script was written with the use of Python programming language. Step I illustrates the fragment of the DNA sequence and matrix with the size 4*4. The first nucleotide of the sequence is A, the corresponding score in the matrix to the letter A is 0.1; next is again A with a score 0.4 and etc. All collected scores have to be multiplied together, consequently a total value of the step I is 0.0024. Matrix shifts to the next nucleotide (Step II) and repeats the same procedure. After all scores have been collected the greatest score has to be selected as the best position for TF binding. In this example, step II gives the highest score, meaning that TF has the highest likelihood of binding to the part of sequence starting from the second base of this sequence.

5.3 P value Concept

In the thesis work the p value concept, the concept of result significance was used. The null hypothesis (H_0) of this work was that TF-match score is likely to occur by chance. Then the alternative hypothesis (H) stated that TF-match score doesn't occur by chance. Hence, a null hypothesis was tested against an alternative. The p value was computed and sorted based on established significant level (α=0.01). The provided data was described with to controversial statements:

- H_0, variant data hasn't pathogenic properties
- H, variant data has pathogenic properties

- P value < α, shows strong evidence against the null hypothesis, meaning all the data that fulfil this condition is actually pathogenic.
- Table 10.1 and Table 10.2 show final results which have deleterious properties.

6 Overall Procedure

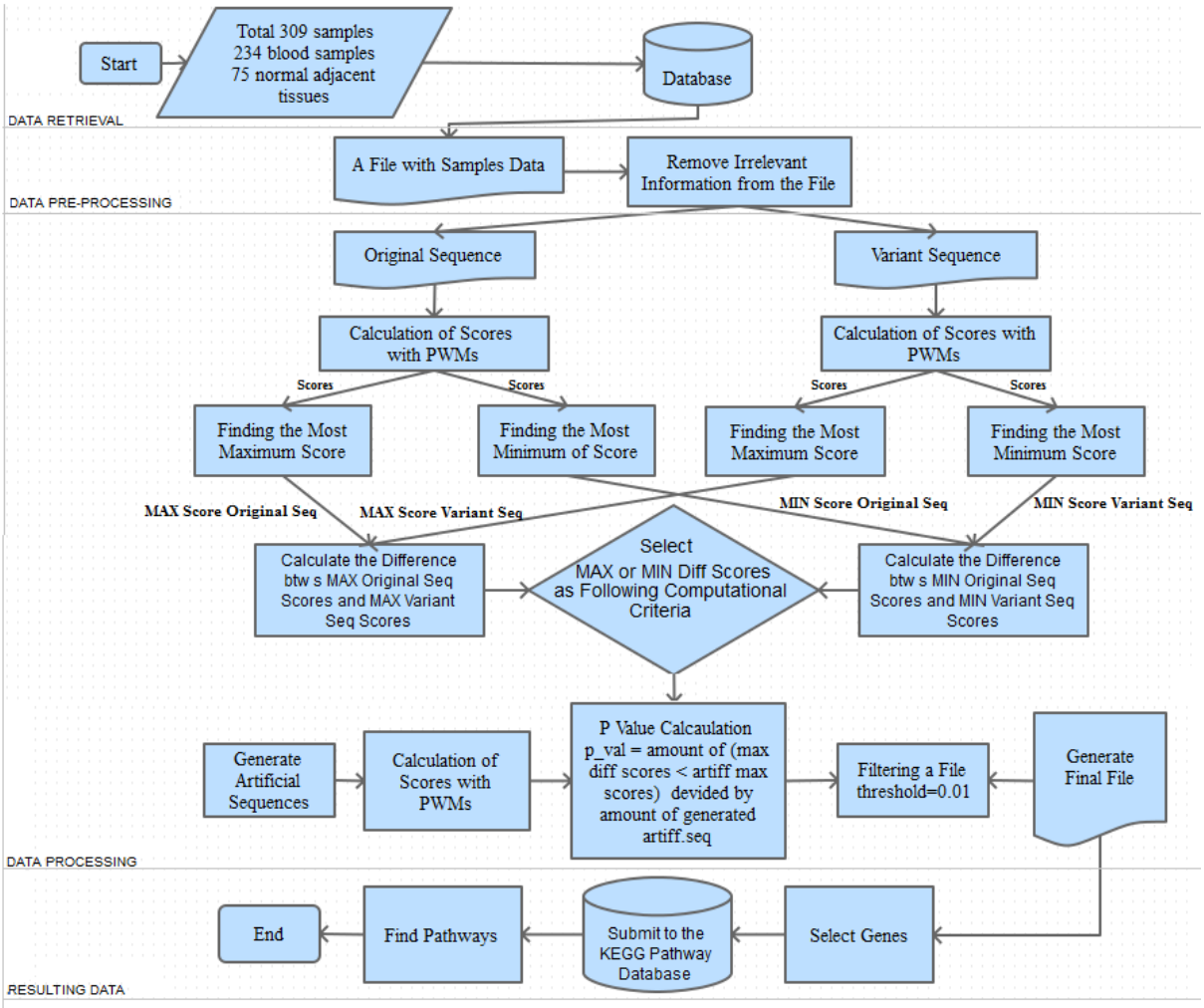

Figure 6.1 Schematic presentation of the computational part workflow

Figure 6.1 shows the flow of the computational part, which starts from extracting, required samples from blood and adjacent tissues. These data is stored and supplied as a file for computational analysis. The computational part was divided into four main parts: data pre-processing, data processing, generating final data and evaluation of results. The detailed description of each step will be described below.

6.1 TCGA Data Source

The data which was provided for the computational part of this thesis derived from patients, as the total amount of 309 samples. There were 234 samples derived from the white blood cells as the source of human DNA information. The rest 75 were extracted from an adjacent normal tissue. These data was functionally annotated with ANNOVAR software. There are 1887215 rows and 41 columns in the ANNOVAR annotated file. Table 6.1 represents the first eight columns and seven rows of this file as an example of the provided information.

Table 6.1 A part of the file with data annotated by ANNOVAR

Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	GeneDetail.refGene
1	10109	10109	A	T	intergenic	NONE,DDX11L1	dist=NONE;dist=1765
1	10177	10177	A	C	intergenic	NONE,DDX11L1	dist=NONE;dist=1697
1	10180	10180	T	C	intergenic	NONE,DDX11L1	dist=NONE;dist=1694
1	10234	10234	C	T	intergenic	NONE,DDX11L1	dist=NONE;dist=1640
1	10235	10235	T	A	intergenic	NONE,DDX11L1	dist=NONE;dist=1639
1	10248	10248	A	T	intergenic	NONE,DDX11L1	dist=NONE;dist=1626
1	10250	10250	A	C	intergenic	NONE,DDX11L1	dist=NONE;dist=1624

Additionally to the information described above the file with PWMs was provided. The information of corresponding PWMs was retrieved from JASPAR and UniPROBE repositories. The file contains the name of the TF at own row and matrix represented as single nucleotide probabilities of size 4^{x} L, where L is the length of a matrix. The length of matrices varies from 6 to 30 across the file.

Table 6.2 A part of the file which contains PWMs

YY1	6					
0.352941	0.01	0.01	1	0.01	0.176471	
0.058824	0.941176	1	0.01	0.01	0.470588	
0.411765	0.01	0.01	0.01	0.01	0.176471	
0.176471	0.058824	0.01	0.01	1	0.176471	
IRF1	2					
0.066667	0.001	0.001	0.733333	0.001	0.001	0.001
0.001	0.066667	1	0.001	0.533333	0.001	0.001
0.001	0.001	0.001	0.001	0.266667	0.001	0.001
0.933333	0.933333	0.001	0.266667	0.2	1	1
GATA2	4					
0.245283	0.01	0.981132	0.01			
0.245283	0.09434	0.01	0.01			
0.339623	0.90566	0.018868	0.01			
0.169811	0.01	0.01	1			

Table 6.2 gives a visual example of the provided file which contains PWMs. There are three examples of matrices; fist row shows TFs name, and the next four rows represent the matrix's row. For example, YY1_6 is TF with 4*6 size matrix; the second is the TF IRF1_2 with 4*7 size matrix; and the last matrix contains the TF GATA2_4 with 4*4 size matrix.

6.2 Data Preparation

6.2.1 Filtering Based on 'exonic' and C Score

Script 1: C-Script.py

```
script FilterExonicAndCScore (inputfile) return outputfile

 input: inputfile, a file with a table (e.g. coordinates, chromosome), where each
        row contains values. Data annotated by ANNOVAR

 output: outputfile, a file that contains the variant data with the columns
         and rows filtered by exonic and C Score


 table = READ-DATA-FROM-FILE (inputfile)

 tmp-table = SELECT-FROM-TABLE-COLUMNS-WITH-VALUES (table, "Func.refGene",
    "exonic")

 tmp-table = SELECT-FROM-TABLE-COLUMNS-WITH-VALUES (table, "cadd_phred", values
    ≥ 20)

 WRITE FILE (tmp-table)
```

Description: The input file contains various variant features annotated by ANNOVAR tool. The output file contains a table with the columns chr, start, end, ref, alt, func.refgene, gene.refgene, genedetail, refgene, exonicfunc.refgene, change.refgene, snp138, cadd.

6.2.2 Subtract One

Script 2: One-Script.py

```
script SubtractOneFromStartCoordinate  (inputfile) return outputfile

 input: inputfile, a file with a table, where values are sorted based on C score
        ≥ 20 and exonic annotation

 output: outputfile, a file that contains variant data where start coordinate
         has value minus one


 table = READ-DATA-FROM-FILE (inputfile)

 tmp-table=SELECT-FROM-TABLE-COLUMN-WITH-VALUES (table, "Start", * )

 tmp-table=SUBTRACT-FROM-COLUMN-A-VALUE (table, "Start", -1)

 WRITE FILE (tmp-table)
```

Description: The input file contains a table with the columns chr, start, end, ref, alt, func.refgene, gene.refgene, genedetail, refgene, exonicfunc.refgene, change.refgene, snp138, cadd. The output file contains a table, where from values of the column 'start' one is subtracted.

6.2.3 BEDTools Functions

BEDTools
BEDTools (*coordinstart_value_minus_one,* tf_sites-encode, genome_inform) **return** *outputfile* **input:** *coordinstart_value_minus_one,* a file that contains the variant data where start coordinated has value minus one*;* *tf_sites_encode,* a file with TF sites from ENCODE database*;* *genome_info,* a file which contains a sequence information in FASTA format **output:** *outputfile*, a FASTA file bedtools INTERSECT –a *coordinstart_value_minus_one* –b *tf_sites_encode* –wo > tf_sites-encode bedtools SLOP –i *coordinstart_value_minus_one* –g *genome_info* –b 25 > *file_seq_extension* bedtools GETFASTA –fi *genome_info* –bed *file_seq_extension* –f > *outputfile*
Description: There are three input files: 1) *coordinstart_value_minus_one* that contains various variant information in BED format, but the main features are chromosome, start coordinated has value minus one, reference and altered nucleotides, TF; 2) *tf_sites-encode* (BED format) contains information about TF sites length from ENCODE database ; 3) *genome_inform* a file in FASTA format with sequence information for homo sapiens. The output file is in FASTA format, contains the information about original sequence, chromosome, start and end coordinates

6.2.4 Generate Variant Sequence

Script 3: VariantSequence.py
script GenerateVariantSequence (*file_seq_extension, file_orig_seq*) **return** *outputfile* **input:** *file_seq_extension, a file that* contains sequence extension information; *file_orig_seq, a file that* contains variant sequence information **output:** *outputfile*, a file that contains variant sequence *list_of_orig_seq* = READ-DATA-FROM-FILE (*file_orig_seq*) *list_of_alt_nucl* = READ-DATA-FROM-FILE (*file_seq_extension*) list_of_variant_seq = [] FOR x IN list_of_orig_seq *replace_nucl* = GET-ALTER-NUCLEOTIDES (*list_of_alt_nucl[x]*) *chromosome* = GET-CHROMOSOME (*list_of_orig_seq[x]*) APPEND-TO-LIST (*list_of_variant_seq, chromosome*) *new_seq* = REPLACE-STRING-WITH-CHAR-AT-POSITION (*list_of_orig_seq*[x], *replacementchar*, 26) APPEND_TO_LIST (*list_of_variant_seq, new_seq*) WRITE_FILE (*list_of_variant_seq*)

Description: The input file *file_seq_extension* (BED format) contains a table with the columns chromosome, start and end coordinates, reference and altered nucleotides, TF. The input file *file_orig_seq* is FASTA format file, contains chromosomal information, start and end coordinates, and original sequence. The output file contains a variant sequence, chromosome, start and end coordinates.
GET-ALTER-NUCLEOTIDES: get the information from the column with altered nucleotides
GET-CHROMOSOME: get the information from the "chr" column about chromosomal position
REPLACE-STRING-WITH-CHAR-AT-POSITION: substitute 26^{th} position of a sequence string by altered nucleotide

6.3 Data Processing

6.3.1 Calculating Scores with PWMs

Script 4: ComputeScoresWithPWMforOriginalAndVarinatSequences.py

Functions and Declarations of the Script

Note: NumPy package was uploaded for this script. Utilize of this package allows to convert matrix into an array

`function PWM_San (input: sequence, output: scores)`

Description: PWMs moves by a single nucleotide along a sequence, a binding score is collected at each position of a nucleotide. When matrix reaches the end of a sequence recorded scores at each nucleotide are multiplied. After matrix scans complete sequence the max score is selected as the best binding score.

`function ConvStrToFloat (input: strings, output: float values)`

Description: the function converts strings into float numbers

`function ConvStrToInteg (input: strings, output: integers)`

Description: the function converts strings into float integers

Description of the Script

```
script ComputeScoreWithPWMconcept  (file_orig_seq, file_pwm) return outputfile
 input: file_orig_seq, contains an original sequence;
        file_pwm, contains Position Weight Matrices (PWMs)
 output: outputfile, a file that contains scores information for the original
         sequence

 list_of_original_seq = READ-DATA-FROM-FILE (file_orig_seq)
 list_pwm = READ-DATA-FROM-FILE (file_pwm)
 list_of_convert_orig_seq = CONVERT-SEQCHAR-INTO-DIGITS (list_of_original_seq)
 list_of_tfs = COLLECT-TF-NAMES-FROM-LIST (list_pwm)
 array_matrices = CONVERT-MATRICES-INTO-ARRAY (list_pwm)
 list_tf_matrix = ZIP-DATA-OF-LISTS (list_of_tfs, array_matrices)
```

```
dictin_tf_matr = GET-KEY-TF-VALUE-ARRAY (list_tf_matrix)

orig_seq_scores_dictionary = {}

FOR x IN list_of_convert_orig_seq

   one_nucleotide = ConvStrToFloat (GET-SEQUENCE( x ))

   FOR key, value IN dictin_tf_matr

          score = PWM_San (one_nucleotide, key, value)

          WRITE FILE (score)
```

Description: The input file *file_orig_seq* contains a table with the columns original sequence, chromosome, start and end coordinates, reference and altered nucleotides, gene name. The input file *file_pwm* contains information about TFs and matrices. The output file contains a table with original sequence, TF name, score, start and end binding sites.
CONVERT-SEQCHAR-INTO-DIGITS: a sequence is representation of nucleotides, which are converted into digits A-1,C-2,G-3,T-4
COLLECT-TF-NAMES-FROM-LIST: takes a list as input, and collects TF names
CONVERT-MATRICES-INTO-ARRAY: takes matrices from a file and converts them into array
GET-KEY-TF-VALUE-ARRAY: takes a dictionary as input and assigns TF as key, and array representation of matrix as value
FOR x IN: takes a string of sequence and converts each character into float number
FOR key, value IN: scans sequence with PWM_Scan function

*The same process runs for variant sequences

6.3.2 Removing Similar Scores

Script 5: RemoveSimilarScore.py

```
script RemoveSimilarScoresBetweenOriginalAndVariantSeq (orig_seq_scores,
       variant_seq_scores) return outputfile

 input: orig_seq_scores, contains original sequence scores;

        variant_seq_scores, contains variant sequence scores

 output: outputfile, a file that contains unique scores for both original and
         variant sequences


 line_orig_seq = READ-DATA-FROM-FILE (orig_seq_scores)

 diff_orig_seq_scores = GET-ORIGSEQ-SCORES (line_orig_seq)

 line_variant_seq = READ-DATA-FROM-FILE (variant_seq_scores)

 IF TF-IS-NOT-EQUAL(line_orig_seq, line_variant_seq) AND SCORE-IS-NOT-EQUAL
   (line_orig_seq, line_variant_seq)

     score_diff = FIND-DIFFERENCE-BETWEEN-SCORES (line_orig_seq, line_variant_seq)

     normalize_score = score_diff / diff_orig_seq_scores

     WRITE-FILE (score_diff, normalize_score)
```

Description: The input file *orig_seq_scores* contains an original sequence, TF name, score, start and end binding sites. The file *variant_seq_scores* contains variant sequence, TF name, score, start and end binding sites. The output file contains original and variant sequences with unique scores, chromosome, start and end coordinates.
FIND-DIFFERENCES-BETWEEN-SCORES: score from original sequence – score from variant sequence
normalize_score: score difference divided by original sequence score

6.3.3 Finding the Most Minimum and the Most Maximum Scores

Script 6: FindMaxMinScore.py

```
script FindTheMostMaxAndMinScores (unique_orig_variant_seq_scores) return
       outputfile
 input: unique_orig_variant_seq_scores, contains unique scores for original
        sequence and variant sequence
 output: outputfile, a file that contains the most maximum and minimum scores for
         original sequence and variant sequence


  orig_var_data_row = READ-DATA-FROM-FILE (unique_orig_variant_seq_scores)
  first_sequence = GET-SEQUENCE-FROM-ROW-FROM-TABLE (orig_var_data_in_one_row, 0)
  current_pos = 0
  WHILE  current_pos <= LEN (orig_var_data_row)
   min_score = 1000
   max_score = 0
   current_sequence = GET-SEQUENCE-FROM-ROW-FROM-TABLE (orig_var_data_row,
    current_pos)
   WHILE (first_sequence == current_sequence)
     IF GET-SCORE (current_sequence) < min_score
              THEN min_score = GET-SCORE (first_sequence)
     IF GET-SCORE (current_sequence) > max_score
              THEN max_score = GET-SCORE (first_sequence)
     current_pos = current_pos + 1
     IF current_pos <= LEN (orig_var_data_row)
        current_sequence = GET-SEQUENCE-ON-POSITION-FROM-TABLE
        (orig_var_data_row, current_pos)
      ELSE
        WRITE-FILE (first_sequence, mins_score, max_score)
        EXIT
   WRITE-FILE (first_sequence, mins_score, max_score)
   first_sequence = current_sequence
```

Description: The input file *unique_orig_variant_seq_scores* contains an original sequence, TF name, original sequence score, start and end binding sites, differences in scores between original and variant sequences, normalization score, variant sequence, TF name, variant sequence score, start and end binding sites. The output file contains original sequence, TF name, maximum value of original sequence score, start and end binding sites, differences in scores between original and variant sequences, normalization score, variant sequence, TF name, maximum value of variant sequence score, start and end binding sites AND original sequence, TF name, minimum value of original sequence score, start and end binding sites, differences in scores between original and variant sequences, normalization score, variant sequence, TF name, minimum value of variant sequence score, start and end binding sites.

6.4 Generating a Final File

6.4.1 Generating Artificial Sequences, Computing Scores with PWMs, Computing P Value

Script 7: GenerateArtificialSequenceComputScoresPvalue.py

Functions and Declarations of the Script

functions *ConvStrToFloat(*input: **strings**, output: **float** values)

Description: the function converts strings into float numbers

```
script GenerateArtificialSeqComputeScoreWithPWMconceptAndPvalue
       (diff_max_score_criteria, file_pwm) return outputfile

 input: max_score_criteria, contains the most maximum scores for original and
        variant sequences;
        file_pwm, contains PWMs information

 output: outputfile, a file that contains p values


 list_of_tfs = COLLECT-TF-FROM-FILE (file_pwm)

 len_matrices = COLLECT-MATRICES-FROM-FILE (file_pwm)

 initial_number_of_artific_seq = 888

 list_of_scores = []

 artif_seq_score = CREATE-ARTIFSEQ-WITH-LENGTH-EQUALS-LENGTH-OF-MATRICES-AND-
  COMPUTE-SCORE (len_matrices)

 APPEND-TO-LIST artif_seq_scores (list_of_scores)

 dictin_tf_matr = KEY-TF-VALUE-SCORES (list_of_tfs, array_matrices)

 values = GET-VALUES-FROM-DICT dictin_tf_matr []

 FOR x IN (max_score_criteria)

   new_seq = REPLACE-DIGITS-WITH-CHAR ( x )

   p_value = HOW-MANY-TIMES (diff_max_score_criteria < values)/ initial_number_seq

   WRITE FILE (p_value, new_seq)
```

Description: The input file *diff_max_score_criteria* contains a table with the columns original sequence, TF name, the most maximum value of original sequence score, start and end binding sites, differences in scores between original and variant sequences, normalization score, variant sequence, TF name, the most maximum value of variant sequence score, start and end binding sites. The input file *file*_pwm contains information about TFs and matrices. The output file contains a table with original sequence, start and end binding sites, TF name, p value, variant sequence, start and end binding sites.

CREATE-ARTIFSEQ-WITH-LENGTH-EQUALS-LENGTH-OF-MATRICES-AND-COMPUTE-SCORE: generate one artificial sequence with the length equals to the length of taken matrix and calculate the score at the same time. As length of artificial sequence is equal to the length of the matrix, for each character of the sequence the relevant score is found from the matrix, consequently these scores are multiplied.

KEY-TF-VALUE-SCORES: create an dictionary, where TF names represented as key, and value as computed with PWMs concept artificial scores (the amount of score equals to amount of generated artificial sequences, 888)

REPLACE_DIGITS_WITH_CHAR: nucleotide sequence was converted into digits, not the task is to converted back into nucleotides 1-A, 2-C, 3-G, 4-T

p_value: calculate the total amount of difference maximum scores that are less than scores of artificially generates sequences divided by amount of generated sequences (888)

6.4.2 Sorting Based on Cut-Off Value

Script 8: SortBasedOnCuttOff.py

```
script SortBasedOnCutOff  (p_value_file) return outputfile

 input: p_value_file, contains the p value found with selected criteria (maximum or
        minimum difference);
 output: outputfile, a file that contains scores less than p value 0.001


 table = READ-DATA-FROM-FILE (p_value_file)

 tmp-table = SELECT-COLUMN-WITH-PVALUES-FROM-TABLE (table, "p values")

 threshold = 0.001

 sorted_table = SORT-PVALUE-COLUMN-BASED-ON-THRESHOLD (tmp-table, threshold)

 WRITE FILE (sorted_table)
```

Description: The input file *p_value_file* contains original sequence, start and end binding sites coordinates, TF name, p value, variant sequence, start and end binding sites. The output file contains the same values as input file but the p values scores are less than 0.001.

6.4.3 Generating a Final File

Script 9: GeneratingFinalFile.py

```
script GenerateresultingFile  (threshold_sorted_p_value_file, reference_sequence,
       Cscore_greater20 ) return outputfile

 input: p_value_file, contains the p value found based on max criteria (or
        minimum);
 output: outputfile, resulting file


 list_orig_seq = READ-DATA-FROM-FILE (reference_sequence)

 list_orig_seq_info = READ-DATA-FROM-FILE (threshold_sorted_p_value_file)

 matched_info = []

 FOR x IN (list_orig_seq_info)

    FOR y IN (list_orig_seq)

        IF  list_orig_seq_info[x] == list_orig_seq [y]

            char_seq = CONVERT-DIGITS-TO-CHAR (list_orig_seq_info [x])

            APPEND-TO-LIST (y)

 c_score = ADD-CSCORE (Cscore_greater20)

 WRITE FILE (matched_info, char_seq, c_score)
```

Description: The input file *p_value_file* contains original sequence, start and end binding sites coordinates, TF name, p value, variant sequence, start and end binding sites. The output file contains resulting information about original sequence, chromosome, start coordinate, reference and altered nucleotides, gene name, TF name, p value, C score and criteria that used to find this information (either maximum or minimum score difference)

reference_sequence is a file *outputfile* created in the beginning of computational part Script 3: BEDTools

6.5 Data Evaluation

A final file was generated and contained information of original sequence, chromosome, start coordinates, reference nucleotide and altered nucleotide, gene name, TF, p value, C score and maximum difference. As it was mentioned maximum and minimum difference criteria were used in parallel that gave the opportunity to compare final results. The gene data was selected and uploaded into free available KEGG database and the resulting pathways were evaluated.

- select genes from the final file
- submit to the KEGG database
- evaluate the results

7 Results and Discussion

The main goal of the thesis was to estimate pathogenicity of human genetic variants by comparing the results produced by the CADD tool and statistical approaches. First of all, the CADD tool would produce variants annotated by C score. As the second main step, the sorted out data based on C score was processed with statistical approaches to reduce amount of false positive results. Eventually the output of both techniques was compared.

The computational part of the thesis was run based on two criteria: the maximum and minimum difference score criteria in parallel to see the differences or similarities in final results. The difference between scores was found by subtracting a score found with PWM of original sequence from a score found with PWM of variant sequence. The resulting files are different in data size such that the data computed based on minimum difference score produces wider range of genes (16873 genes) than based on maximum difference score (6752 genes).

Majority of different processes in our body are derived, controlled and regulated by proteins that are manufactured by genes. Accordingly, pathways allocations of the resulting genes were not surprising. The majority genes were involved into the metabolic pathways, the process that enables a cell to keep living, growing and dividing due to the set of chemical reactions. The second place was related to the cancer pathways and the last to neuroactive ligand-receptor interaction.

Hence, 6752 genes were submitted to the KEGG pathway search tool from the resulting file based on maximum difference scores criteria. Consequently, 156 genes represent metabolic pathways, 62 genes characterize pathways in cancer, and 52 in neuroactive ligand-receptor interaction. There were eight genes such as, CREB3, EGF, ERBB2, FGFR1, FOXO1, IKBKG, LEF1, RB1 indicated under prostate cancer pathway by KEGG database.

In a like manner, the same outcome pathways derived from the file based on minimum difference scores criteria. However, the amount of genes three times exceeded comparing to the previous file. There were 272 genes that represent metabolic pathways, 116 genes characterized pathways in cancer, and 102 in neuroactive ligand-receptor interaction pathways. The KEGG database indicated 21 genes under prostate cancer pathway, such as ARAF, CASP9, CCND1, CDKN1B, CREB3, CREB5, E2F2, EGF, EP300, ERBB2, FGFR1, FOXO1, IGF1R, IKBKB, IKBKG, LEF1, PDGFB, PDGFRA, PDGFRB, RB1, TCF7.

In the beginning of the computational part before using the statistical approaches the provided file was cleaned out for the data with C score greater than 20. Table 9.2 illustrates the pathogenic variants; where the data was sorted out based on C score greater than 20. The table represents the chromosomal position, start coordinate, reference and alternate nucleotide, the genome region, gene, type of mutation, dbSNP rs identifiers and C score. The result of this table will be compared to the final result after statistical calculations.

Figure 7.1 illustrates the most frequent genes with C score greater than 20. The set of these genes with predicted harmful properties was eventually submitted into KEGG database and related pathways were extracted.

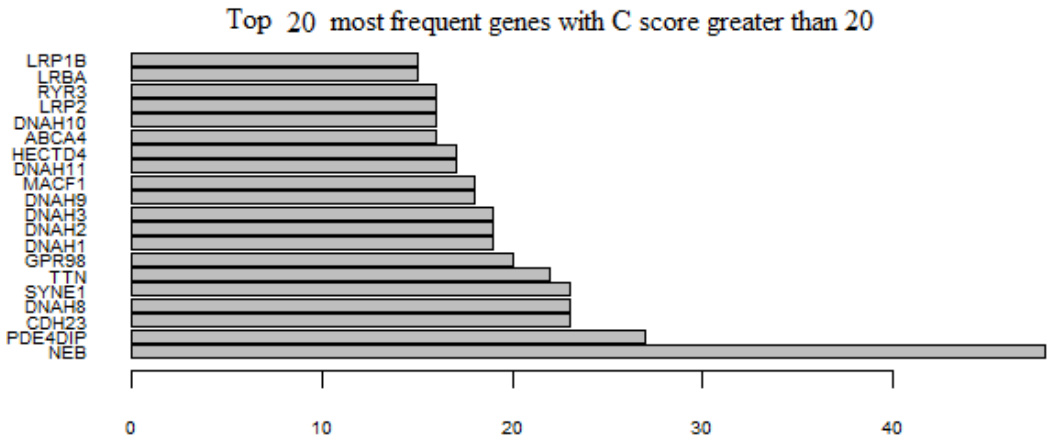

Figure 7.1 Most frequent genes with C score greater than 20

The CADD tool predicted only 0,88 % of variants from provided data as deleterious, the same result was achieved by using statistical approach. Consequently only 0.88% from the given data represents deleterious variants.

The CADD tool represents in one score (the C score) all various characteristics that associated with pathogenic variants. That was one of the main interest to check if a set of genes with C score greater than 20 matched the set of genes in the resulting file. Surprisingly, the KEGG database shown 21 genes that lead to the prostate cancer such as ARAF, CASP9, CCND1, CDKN1B, CREB3, CREB5, E2F2, EGF, EP300, ERBB2, FGFR1, FOXO1, IGF1R, IKBKB, IKBKG, LEF1, PDGFB, PDGFRA, PDGFRB, RB1, TCF7 with 100 % match. Consequently, these results confirm that CADD tool revealed correctly genes that lead to the prostate cancer comparing to the results which were elaborated with statistical methods to exclude false positive rates. As it was discussed in the section 4.4 the CADD tool combines a lot of different features (63 distinct annotations) related to identification of deleteriousness in one score, and this makes the output extremely accurate. On the other hand the provided data was relatively small in size, compared to the resulting variant-by-annotation matrix contained 29.4 million variants (half observed, half simulated).

There is 100% match between genes that were accomplished by three different criteria. For further analysis the wider range of information was taken: sequence, chromosomal, start, gene name, TF, p value, C score and minimum difference score information were extracted (Table 9.1 Linked data from the final file with minimum criteria and ANNOVAR file Table 9.2 Filtered data based on C score greater than 20). The C score column of the table was sorted in descending way, and as the result three the most deleterious (according to the C score) genes, such as PDGFRA, CREB5, FGFR1 were selected (Table 7.1).

The C score indicates the pathogenic properties of a variant, higher C score more harmful variant is predicted to be. Due to the specificity of computational techniques the p value is equal to zero than in reality it is just a score that is very close to the zero. The higher value of C score was correlated to the zero value of the p value; there was no strong correlation to the minimum difference score, however it was still relatively small in range of -0.001 to -0.04.

Table 7.1 Top three resulting genes with the highest C score

Original Sequence	Chr/ Start-End	Ref	Alt	Gene	TF	P	Cadd	Min Diff Score
AGAGACTGAGC GCTGACAGTGG CTACATCATTCC TCTGCCTGACAT TGACC	chr4: 55156627-55156678	C	A	PDGFRA	CREB1_5	0	43	-0.045464496
ACAATACAGCC ACCCCAGCCCA CAGGGGGGCGC CGGCGAAGGGT GGTAGAC	chr7: 28848839-28848890	G	C	CREB5	USF2_1	0	36	-0.01507
AGGTCTGACAA GTCTTTCTCTGT TGCGTCCGCTTT AAAGAACACGT TGAGA	8chr8: 38274903-38274954	C	T	FGFR1	RFX1	0	36	-0.0017136

One of the top three genes was the PDGFA gene with identified TF CREB1. PDGFA is a subunit of a typical cancer gene platelet-derived growth factor (PDGF) ("GeneCards - Human Genes | Gene Database | Gene Search," n.d.). The PDGF and its isoforms PDGFA, PDGFB heterodimer (all three present in the final result) are responsible for cell proliferation, cellular differentiation, cell growth, development (Heldin, 2013). The resulting data showed the non-synonymous single nucleotide variation. Overexpression or mutation events of this gene might lead to the tumour cell growth (Heldin, 2013). Below Figure 7.2 shows transcription starting site of PDGFA, and four possible binding sites of TF CREB1 ("Sample to Insight - QIAGEN," n.d.).

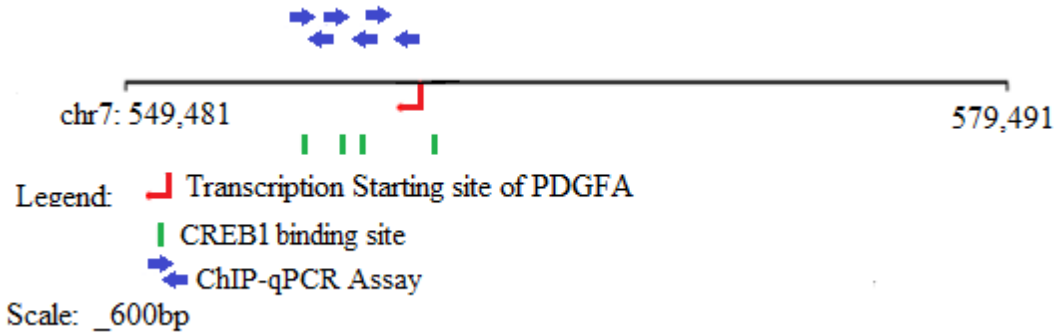

Figure 7.2 Transcription starting site of the PDGFA gene and possible binding sites of the TF CREB1 ("Sample to Insight - QIAGEN," n.d.)

The transcriptional factor CREB1 belongs to the leucine zipper family of DNA binding proteins. Its function is to bind to certain DNA sequences such as *cyclic adenosine monophosphate* (cAMP) response elements, as a result it regulates a gene expression (a transcription of the downstream genes is increased or decreased) ("GeneCards - Human Genes | Gene Database | Gene Search," n.d.). Overall, the properties of TF CREB1 that impact on the gene expression, might be altered due to mutation events in the gene PDGF or its isoforms, consequently leading to the over activity of PDGFA and eventually tumour cells' growth. Moreover, it was proved by microarray analyses that PDGFRβ mRNA expression relapse prostate cancer (Heldin, 2013).

The second in the range was gene CREB5 and TF USF2. The CREB5 is a protein-coding gene that belongs to cyclic AMP-responsive element-binding family, activates transcription ("GeneCards - Human Genes | Gene Database | Gene Search," n.d.). Figure 7.3 shows four possible transcription starting sites and only one binding site of TF USF2 ("Sample to Insight - QIAGEN," n.d.).

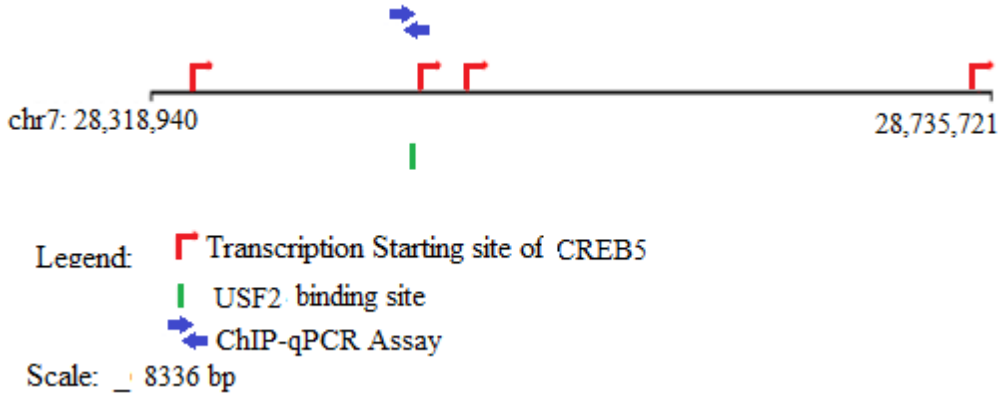

Figure 7.3 Transcription starting site of the gene CREB5 and binding site of the TF USF2 ("Sample to Insight - QIAGEN," n.d.)

The USF2 is transcriptional factor that binds to a symmetrical DNA-sequence. The USF2 gene provides instructions for making a protein called USF2, it impacts on cellular growth and proliferation. It involved in the series of events of transferring genetic information from DNA to mRNA by DNA-directed RNA polymerase. One of its functions is to act as cellular TF.

N Chen and others studied the role of USF2 in prostate tumorigenesis. It was found that one possible way of developing cancer by USF2 is to regulate many cancer- and proliferation-

associated genes (Cox-2, BRCA2, p53 and etc). This happens due to the correlation of its function to transcriptional activity (Chen et al., 2006).

Finally, the third gene FGFR1 and TF RFX1 were identified as genes related to the cancer pathway. The FGFR1, is also called fibroblast growth factor receptor 1 is involved into cell division, growth, maturation formation of blood vessels.

According to Yang F FGFR1 gene expressed in prostate carcinoma cells is not specific for this type of genes, mostly it is not expressed in epithelial cells. Even though the role and mechanism of this gene is not completely understood, Yang and the team suggest that FGFR1 signalling is a key regulator of prostate cancer proliferation, histopathological phenotype, and cancer progression to metastasis (Yang et al., 2013).

TF RFX1 is a member of the regulatory factor X (RFX) gene family. It has wide range of functions starting from response of DNA damage ending in ciliary gene regulation (Min et al., 2014).

In addition the final data was linked to the initial results that were extracted from ANNOVAR. The ANNOVAR tool provides information about the functional consequences of the variants by annotating them. The resulting data is shown in the Appendix (Table 9.1 and Table 9.2). The fourth column shows type of mutation characterization for a variant. There is a vast amount of data represented mostly as stopgain in the beginning of the table with the highest C score. Stopgain defines by ANNOVAR as non-synonymous SNV or frameshift insertion/deletion mutations that lead to termination of translation (stop codon). When the C score turns to the smaller values variants are defined mostly as non-synonymous SNV. In this thesis only single nucleotide substitution was observed. Since the core role of the stopgain is to terminate translation, this fact makes it obvious that a final protein would be synthesized incorrectly.

There is a huge amount of genes present in a human organism, which have various duties. All these genes can be grouped into three groups regarding their contribution to cancer.

The main group is tumour suppression genes, where they regulate cells growth, by monitoring the speed of division, repairing mismatched DNA and controlling a cell live time. Mutations in these genes will affect this process dramatically turning the process from positive into tumour growing. The typical member of this group is BRCA1 and BRCA2, which are also observed in the final data Table 7.2.

Table 7.2 BRCA2 and BRCA1 from resulting file with minimum criteria

Chr	Start	Gene	Mutations	TF	P value	C score	Min Diff
13	32972626	BRCA2	Stopgain	EN1_EN2_1	0	51	-0.033382742
17	41234556	BRCA1	stopgain	NFIC_1	0	37	-0.040655136
13	32954181	BRCA2	nonsynonymous SNV	ZNF354C_1	0	36	-0.005737616
13	32953550	BRCA2	nonsynonymous SNV	HNF4A_5	0	24.9	-0.011436927
17	41203095	BRCA1	nonsynonymous SNV	NFYA_2	0	22.5	-0.13351328
13	32972525	BRCA2	nonsynonymous SNV	UBP1::TFCP2_1	0	21.9	-0.037197775
17	41219631	BRCA1	nonsynonymous SNV	HLX_2	0	20.9	-0.009312052
17	41249297	BRCA1	nonsynonymous SNV	HSF1_HSF2_1	0	20.3	-0.999

Two top genes BRCA2 and BRCA1 with the highest C score 51 and 37 respectively are considered as deleterious and they are also defined by ANNOVAR tool as stopgain, meaning the mutations in these genes disrupted the production of a correct protein by stop codon process. The rest of these two genes are defined as non-synonymous SNV that is clearly the process of nucleotide substitution that is not considered as harmful.

The other category of genes that contributes to cancer is oncogenes that turn healthy cells into cancerous cell. One of the typical genes associated with this group is WNT. Table 7.3 represents 19 genes that belong to WNT gene family. There are two genes WNT5A and WNT11 that have the highest C score. Mutations in these genes lead to uncontrolled cells growth.

Table 7.3 WNT family genes from resulting file with minimum criteria

Chr	Start	Gene	Mutations	TF	P value	C score	Min Diff
3	55508479	WNT5A	stopgain	RUNX1_1	0	38	-0.001646881
11	75902750	WNT11	stopgain	DEAF1_4	0	38	-0.038779263
7	120978983	WNT16	nonsynonymous SNV	GABPA_1	0	36	-0.059504176
1	113058989	WNT2B	stopgain	EVX1_EVX2_2	0	35	-0.011559067
2	219736411	WNT6	nonsynonymous SNV	MZF1_3	0	35	-0.028536945
2	219754966	WNT10A	nonsynonymous SNV	E2F1_	0	34	-0.4809375
2	219747090	WNT10A	stopgain	NR3C1_3	0	33	-0.018505368

22	46318765	WNT7B	nonsynonymous SNV	MTF1_3	0	33	-0.012622572
12	49363980	WNT10B	nonsynonymous SNV	HNF4A_6	0	32	-0.009207
17	44847372	WNT3	nonsynonymous SNV	WT1_1	0	29.7	-0.026367188
2	219755011	WNT10A	nonsynonymous SNV	IRF8_2	0	29.4	-0.034280205
1	113059840	WNT2B	nonsynonymous SNV	E2F1_	0	24.1	-0.187125
3	13896262	WNT7A	nonsynonymous SNV	E2F1_	0	24	-0.002209947
1	22446566	WNT4	nonsynonymous SNV	ZNF354C_1	0	23.3	-0.001610957
17	44952508	WNT9B	nonsynonymous SNV	E2F1::TFDP1_1	0	22.8	-0.1875
12	49374437	WNT1	nonsynonymous SNV	GABPA_1	0	22.7	-0.004284668

The last group is DNA repair genes that correct mistakes during process of cells division. Obviously mutations in these genes can lead to the lack of repair. There are different types of DNA repair where various genes are participating. BRCA1 and WNT are one example of typical for this process gene.

It is important to be aware of certain limitations that can bias the results produced by selected tools. The quality of the experimental data affects weights, used in the PWMs. Moreover, the sensitivity and specificity of the PWMs are at the low level (Gershenzon et al., 2005). The CADD tool also has limitations, for example the accuracy of the C score might be affected by the local mutation rate, background selection, and biased gene conversion parameters. There is a need of the 'gold standard' data, which can help in annotating variants better (Kircher et al., 2014).

8 Conclusion

This section compiles the key-results of the thesis work. The main result of the thesis is the following basic conclusions:

- The set of putative genes was produced as the final file;
- The computational implementation of the PWMs work allowed to compute TFs binding scores, which are essential in predicting the TFBSs;

The first research question, *"Q1: Which variants from provided input data have potential for "dual codon usage?"* led to the set of variants that have potential for 'dual codon usage' and was represented in the final file. It was important to extract and record those variants from provided data that could have potential for 'duons'. Moreover, three types of genes PDGFA, CREB5, FGFR1 with TFs as CREB1, USF2, and RFX1 respectfully were analysed in the section 'Results and Discussion'. Mutations in a gene structure could lead to the production of the same amino acids as they can be encoded by a multiple combination of nucleotides. However, the TF that binds to such region receives a wrong pattern and obviously wrong instructions are given to the expression of a gene. Table 7.1 illustrates that the PDGFRA gene has alteration of nucleotide C into A, where TF CREB1_5 binds to the altered sequence thus represents pathogenic properties.

The second research question, *'Q2: How TFBSs can be predicted by applying quantitative PWMs framework?'* indicates the quantitative approach called PWMs, which is used for computing TFs binding scores and predicting the best binding site. It is essential to know TF binding sites as this process reflects on a gene regulation process.

The third research question *'Q3: What kind of tools can be applied for implementing theoretical concepts to lead the research to the final result?*' was answered that CADD and PWMs were the main sources used to achieve final results. The CADD tool was used as the most significant tool for detecting deleterious variants from provided data. As the major feature of this tool is combination of different characteristics used by other recourses into one score, that makes evaluation of an output significantly easy. This thesis work proved that the computational outcome by CADD tool is very accurate and time efficient, and indeed can be used as the trustful resource for scoring and labelling variants. The comparison of statistical analysis and CADD results of this project proved that C score produced by CADD tool determines the data correctly as pathogenic. On the other hand the perfect match of CADD tool might be explained by a relatively small amount of examined data and possibly perfect match to the variants that were already detected as deleterious by many other resources. The concept of computing TF binding scores was based on PWMs, which was implemented by Python scripting language.

Overall analysis indicates that exome regions carry a disease causing amount of data, which requires thorough investigation and cataloguing. Creation of variants catalogues can help

bioscience in further investigations. It makes the job easier in comparing results to already existing variants catalogues and eliminating irrelevant data. It has also been seen that dual work of codons might lead to the distraction of the instructions that TFs provide than bind to the mutated sequence.

Only 0.88% variants from entire provided data were identified as deleterious. These genes from the resulting data were analysed for capability to develop prostate cancer. The KEGG pathway database supports the view that the potential of these genes can progress into prostate cancer.

9 APPENDIX A

Table 9.1 Linked data from the final file with minimum criteria and ANNOVAR file

Chr	Start	Gene	ExonicFunc.refGene	TF	P value	C score	Min Diff
19	9033237	MUC16	stopgain	NKX3-2_1	0	61	-0.009682629
2	179463948	TTN	stopgain	DEAF1_5	0	60	-0.10699776
2	179473091	TTN	stopgain	LEF1_1	0	60	-0.0082875
7	100389677	ZAN	unknown	NFIC_1	0	59	-0.023013089
7	100392843	ZAN	unknown	CREB1_5	0	59	-0.431372216
1	89729430	GBP5	stopgain	ETS1_6	0	57	-0.002495556
19	7964978	LRRC8E	nonsynonymous SNV	DEAF1_5	0	56	-0.004070077
4	1389433	CRIPAK	stopgain	MYC_1	0	55	-0.196483597
16	67318742	PLEKHG4	stopgain	GABPA_1	0	55	-0.004420431
4	1388436	CRIPAK	stopgain	MTF1_3	0	54	-0.329175
4	1389215	CRIPAK	stopgain	AHR_ARNT_HIF1A_1	0	54	-0.80919
6	38998103	DNAH8	stopgain	HSF1_HSF2_2	0	54	-0.168066747
14	64560092	SYNE2	stopgain	TEAD1_1	0	53	-0.000197554
16	20946773	DNAH3	stopgain	BRCA1_1	0	53	-0.006708936
2	152474966	NEB	stopgain	YY1_6	0	52	-0.0143856
16	20944746	DNAH3	stopgain	BRCA1_1	0	52	-0.002148221

Table 9.2 Filtered data based on C score greater than 20

Chr	Start	Ref	Alt	Func.refGene	Gene	ExonicFunc.refGene	Snp138	C score	Min Diff
19	9033237	G	T	exonic	MUC16	stopgain	NA	61	-0.009682629
2	179463948	G	A	exonic	TTN	stopgain	NA	60	-0.10699776
2	179473091	C	A	exonic	TTN	stopgain	rs79432997	60	-0.0082875
7	100389677	C	T	exonic	ZAN	unknown	rs149104440	59	-0.023013089
7	100392843	T	A	exonic	ZAN	unknown	NA	59	-0.431372216
1	89729430	T	A	exonic	GBP5	stopgain	NA	57	-0.002495556
19	7964978	G	A	exonic	LRRC8E	non-synonymous SNV	rs370440409	56	-0.004070077
4	1389433	C	A	exonic	CRIPAK	stopgain	rs145208075	55	-0.196483597
16	67318742	C	T	exonic	PLEKHG4	stopgain	rs142861229	55	-0.004420431
4	1388436	C	G	exonic	CRIPAK	stopgain	rs367925864	54	-0.329175
4	1389215	C	T	exonic	CRIPAK	stopgain	rs112507956	54	-0.80919
6	38998103	C	T	exonic	DNAH8	stopgain	rs146551804	54	-0.168066747
14	64560092	G	A	exonic	SYNE2	stopgain	rs2781377	53	-0.000197554
16	20946773	C	T	exonic	DNAH3	stopgain	rs144426187	53	-0.006708936
2	152474966	C	T	exonic	NEB	stopgain	NA	52	-0.0143856
16	20944746	C	T	exonic	DNAH3	stopgain	rs377349475	52	-0.002148221

10 APPENDIX B

Table 10.1 Final result, based on max difference criteria

Original Sequence	Chr	Start	Ref	Alt	Gene	TF	P value	C score	Max Diff
CCATATTTATTCTGGGCCATGATT CGGAATACATATTCATGGCCTTCTAGC	2	179463948	G	A	TTN	DEAF1_5	0	60	0.0336
GGGCTTTTCCCCCAGGAGCTAGCT CGGGCAGCCACCCTGGAGAGCCTCCGG	19	7964978	G	A	LRRC8E	DEAF1_5	0	56	0.0594
TGGAGTGCCCGCCTGCTCACACGTG CCCATGTGGAGTGCCTGCCTGCTCAC	4	1389433	C	A	CRIPAK	MYC_1	0	55	0.2673
CCCTGACTTGCCTCCTGCCCACTTCC GAAAGATGTGGGCTCTGGCCACGGG	16	67318742	C	T	PLEKHG4	GABPA_1	0	55	0.4137
GGAGTGCCCGCCTGCTCACACGTGCC GACGTGGAGTGCCCGCCTGCTCACG	4	1389215	C	T	CRIPAK	AHR_ARNT_HIF1A_1	0.0011	54	0.0016
GTAGTCATAAAACAGACCAATGAATG GGATGAAGAAATAGAAAATTTGAAA	14	64560092	G	A	SYNE2	TEAD1_1	0	53	0.2398
TCGGAACCCAGCCAATGCCTTTCATC CATTCAAGGTCTGACTTATACAAAT	2	152474966	C	T	NEB	YY1_6	0	52	0.0275
GCTAAGTCAGTGGGATAGCCCAATGC GAGTGAAGCTGTCAATCTGGAAGCC	8	100832183	C	T	VPS13B	ZFP161_2	0	51	0.0006
AGGACATGAAAATTATGGCAGAAAAG AAGAACCAGATCATACTTATGAACC	3	130187662	G	T	COL6A5	SPI1_1	0	50	0.0088
GTAGCTGCACAAAAGGGGGACAGGCC CACCTTTCCTGCTGTTTTAAGGACT	1	144852390	C	T	PDE4DIP	ZNF354C_1	0	49	0.3746
ACACCTCCCTGAGCTACTCTGGCTGGG GTGGCAGCAGTGGGCACGATGCCC	3	119306449	G	A	ADPRH	CHURC1_1	0	49	0.0115
TTTGGTTGGTCATGAGATGGAAAAAGT AGCCATAGCCAGCACCACACACTC	6	51752011	G	T	PKHD1	ZNF384_1	0	49	0.0154
TGAATGGGTCAATGGAGGTGCCCTTCA GTTCCATAACTTTGTGATGGTGAA	8	110477066	C	T	PKHD1L1	IRF1_2	0	49	0.1702
TGTGACAAACAGAAGTCTTGCATTTGAA GAAGGAAGCCAGAATACAACTAT	11	108183151	G	T	ATM	SPI1_1	0	49	0.0184
AAGAAAGAGGCTTTCAGATTCTAAAG GAAAGAATACATGCGGTGGATTTTT	13	103527930	G	T	BIVM-ERCC5,ERCC5	NFATC2_1	0	49	0.1514
TTGAACTTTTGTTTCCGCCTGTTTCC ATAAAGTACAGATGTCTTCCAGGCC	16	70016361	C	T	PDXDC2P	NFATC1_1	0	49	0.0974

Table 10.2 Final result, based on min difference criteria

Original Sequence	Chr	Start	Ref	Alt	Gene	TF	P value	C score	Min Diff
TTCTTAAGGAGCGTCTCACCTGAGT GAGGCTAGTCTGCAGCCTGAATAGAG	19	9033237	G	T	MUC16	NKX3-2_1	0	61	-0.009683
CCATATTTATTCTGGGCCATGATTC GGAATACATATTCATGGCCTTCTAGC	2	179463948	G	A	TTN	DEAF1_5	0	60	-0.106998
TGTGTACTGTTAACTTCACGTTTTT CAAGCCAGTAACCCAAAATGGGGCTT	2	179473091	C	A	TTN	LEF1_1	0	60	-0.008288
TCCGAATGTAGCCCGGAGCAGCTGG CGAGCAACAGCACCCAGGCCTGTAGG	7	100389677	C	T	ZAN	NFIC_1	0	59	-0.023013
AAAGCCCGTGTCTGCAGAACCCCTG TCAGAATGACGGGCAGTGTCGGGAGC	7	100392843	T	A	ZAN	CREB1_5	0	59	-0.431372
AGAACTAGGGATACCTGTATTCCTT TCCGAGGCTCCCGATAGTACTTTGCC	1	89729430	T	A	GBP5	ETS1_6	0	57	-0.002496
GGGCTTTTCCCCCAGGAGCTAGCTC GGGCAGCCACCCTGGAGAGCCTCCGG	19	7964978	G	A	LRRC8E	DEAF1_5	0	56	-0.00407
TGGAGTGCCCGCCTGCTCACACGTG CCCATGTGGAGTGCCTGCCTGCTCAC	4	1389433	C	A	CRIPAK	MYC_1	0	55	-0.196484
CCCTGACTTGCCTCCTGCCCACTTC CGAAAGATGTGGGCTCTGGCCACGGG	16	67318742	C	T	PLEKHG4	GABPA_1	0	55	-0.00442
TGTCGATGCGGAGTGCCCGCCTGCT CACACATGCCCATGTGGAGTGCCCGC	4	1388436	C	G	CRIPAK	MTF1_3	0	54	-0.329175
GGAGTGCCCGCCTGCTCACACGTGC CGACGTGGAGTGCCCGCCTGCTCACG	4	1389215	C	T	CRIPAK	AHR_ARNT_ HIF1A_1	0	54	-0.80919
GACCTTCATCACTGTGGTATATTTA CGAACAGTGTTGTCCCCGGATCACTG	6	38998103	C	T	DNAH8	HSF1_ HSF2_2	0	54	-0.168067
GTAGTCATAAAACAGACCAATGAAT GGGATGAAGAAATAGAAAATTTGAAA	14	64560092	G	A	SYNE2	TEAD1_1	0	53	-0.000198
CTGTGTGAAGTAGAATCCAGAGATCCA AAATACCACAGGGGGCCCCTTGTC	16	20946773	C	T	DNAH3	BRCA1_1	0	53	-0.006709
TCGGAACCCAGCCAATGCCTTTCATC CATTCAAGGTCTGACTTATACAAAT	2	152474966	C	T	NEB	YY1_6	0	52	-0.014386
CCCCAATCTGCATCGTTTTCCTGTCC CAACGGGCACCTTCTAAGAAGAGCC	16	20944746	C	T	DNAH3	BRCA1_1	0	52	-0.002148
TGTTGGTAGGTTGAGGGCAAATGATG AAGTTCTCAGCTTCTTATAGATTTG	2	21228410	G	T	APOB	GATA2_4	0	51	-0.116028
GCTAAGTCAGTGGGATAGCCCAATGC GAGTGAAGCTGTCAATCTGGAAGCC	8	100832183	C	T	VPS13B	ZFP161_2	0	51	-0.091146
GAATTCTCCTCAGATGACTCCATTTA AAAAATTCAATGAAATTTCTCTTTT	13	32972626	A	T	BRCA2	EN1_EN2_1	0	51	-0.033383

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–9. doi:10.1038/nmeth0410-248

AEpiA :: Australian Epigenetic Alliance. (n.d.). Retrieved February 13, 2015, from http://www.epialliance.org.au/contents/AboutUs/WhatIsEpigenetics.shtml

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). Molecular Biology of the Cell. Garland Science. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK21054/

An Overview of the Human Genome Project. (n.d.). Retrieved February 13, 2015, from http://www.genome.gov/12011238

Atlas of Genetics and Cytogenetics in Oncology and Haematology. (n.d.). Retrieved February 13, 2015, from http://atlasgeneticsoncology.org/

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., … Zhang, J. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Computational Biology*, *9*(11), e1003326. doi:10.1371/journal.pcbi.1003326

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews. Genetics*, *12*(11), 745–55. doi:10.1038/nrg3031

bedtools: a powerful toolset for genome arithmetic. (n.d.). Retrieved February 13, 2015, from http://bedtools.readthedocs.org/en/latest/

Bolsover, S. R., Shephard, E. A., White, H. A., & Hyams, J. S. (2011). *Cell Biology: A Short Course*. John Wiley & Sons. Retrieved from https://books.google.com/books?id=Kt_hL1stQQkC&pgis=1

Box 1 : Exome sequencing as a tool for Mendelian disease gene discovery : Nature Reviews Genetics. (n.d.). Retrieved February 13, 2015, from http://www.nature.com/nrg/journal/v12/n11/box/nrg3031_BX1.html

Bryois, J., Buil, A., Evans, D. M., Kemp, J. P., Montgomery, S. B., Conrad, D. F., … Dermitzakis, E. T. (2014). Cis and trans effects of human genomic variants on gene expression. *PLoS Genetics*, *10*(7), e1004461. doi:10.1371/journal.pgen.1004461

Buske, O. J., Manickaraj, A., Mital, S., Ray, P. N., & Brudno, M. (2013). Identification of deleterious synonymous variants in human genomes. *Bioinformatics (Oxford, England)*, *29*(15), 1843–50. doi:10.1093/bioinformatics/btt308

Chen, N., Szentirmay, M. N., Pawar, S. A., Sirito, M., Wang, J., Wang, Z., … Sawadogo, M. (2006). Tumor-suppression function of transcription factor USF2 in prostate carcinogenesis. *Oncogene*, *25*(4), 579–87. doi:10.1038/sj.onc.1209079

Cheung, K. H., Osier, M. V, Kidd, J. R., Pakstis, A. J., Miller, P. L., & Kidd, K. K. (2000). ALFRED: an allele frequency database for diverse populations and DNA polymorphisms. *Nucleic Acids Research*, *28*(1), 361–3. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102486&tool=pmcentrez&rendertype=abstract

Chun, S., & Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Genome Research*, *19*(9), 1553–61. doi:10.1101/gr.092619.109

Clancy, S. (2008). RNA Transcription by RNA Polymerase: Prokaryotes vs Eukaryotes. Retrieved February 13, 2015, from http://www.nature.com/scitable/topicpage/rna-transcription-by-rna-polymerase-prokaryotes-vs-961

DeWeerdt, S. E. (2004). What's a Genome? Retrieved February 13, 2015, from http://www.genomenewsnetwork.org/resources/whats_a_genome/Chp1_1_1.shtml

Djordjevic, M., Sengupta, A. M., & Shraiman, B. I. (2003). A biophysical approach to transcription factor binding site discovery. *Genome Research*, *13*(11), 2381–90. doi:10.1101/gr.1271603

DNA Is Constantly Changing through the Process of Mutation. (n.d.). Retrieved February 13, 2015, from http://www.nature.com/scitable/topicpage/dna-is-constantly-changing-through-the-process-6524898

Drummond, D. A., & Wilke, C. O. (2009). The evolutionary consequences of erroneous protein synthesis. *Nature Reviews. Genetics*, *10*(10), 715–24. doi:10.1038/nrg2662

Edwards, S. L., Beesley, J., French, J. D., & Dunning, A. M. (2013). Beyond GWASs: illuminating the dark road from association to function. *American Journal of Human Genetics*, *93*(5), 779–97. doi:10.1016/j.ajhg.2013.10.012

GeneCards - Human Genes | Gene Database | Gene Search. (n.d.). Retrieved February 13, 2015, from http://www.genecards.org/

Genome Reference Consortium. (n.d.). Retrieved May 10, 2015, from http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/

Genome-Wide Association Studies Fact Sheet. (n.d.). Retrieved February 13, 2015, from http://www.genome.gov/20019523

Gershenzon, N. I., Stormo, G. D., & Ioshikhes, I. P. (2005). Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. *Nucleic Acids Research*, *33*(7), 2290–301. doi:10.1093/nar/gki519

Glossary. (n.d.). Retrieved April 25, 2015, from http://web.ornl.gov/sci/techresources/Human_Genome/glossary.shtml

Heldin, C.-H. (2013). Targeting the PDGF signaling pathway in tumor treatment. *Cell Communication and Signaling : CCS*, *11*(1), 97. doi:10.1186/1478-811X-11-97

Hume, M. A., Barrera, L. A., Gisselbrecht, S. S., & Bulyk, M. L. (2015). UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, *43*(Database issue), D117–22. doi:10.1093/nar/gku1045

KEGG: Kyoto Encyclopedia of Genes and Genomes. (n.d.). Retrieved February 13, 2015, from http://www.genome.jp/kegg/

Keightley, P. D., & Eyre-Walker, A. (2007). Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*, *177*(4), 2251–61. doi:10.1534/genetics.107.080663

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–5. doi:10.1038/ng.2892

L. Eisenstadt. (2010). What is exome sequencing? | Broad Institute of MIT and Harvard. Retrieved February 13, 2015, from http://www.broadinstitute.org/blog/what-exome-sequencing

Laity, J. H., Lee, B. M., & Wright, P. E. (2001). Zinc finger proteins: new insights into structural and functional diversity. *Current Opinion in Structural Biology*, *11*(1), 39–46. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11179890

Laszlo P. (2009). Protein Evolution (2nd Edition). Retrieved February 13, 2015, from http://library.alibris.com/Protein-Evolution-Laszlo-Patthy/book/5434908

Latchman, D. S. (1997). Transcription factors: an overview. *The International Journal of Biochemistry & Cell Biology*, *29*(12), 1305–12. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9570129

Latchman, D. S. (2008). *Eukaryotic transcription factors*. *Elsevier/Academic Press*. Amsterdam. Retrieved from http://www.ncbi.nlm.nih.gov/nlmcatalog/101463356

Li, M.-X., Kwan, J. S. H., Bao, S.-Y., Yang, W., Ho, S.-L., Song, Y.-Q., & Sham, P. C. (2013). Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genetics*, *9*(1), e1003143. doi:10.1371/journal.pgen.1003143

Li, Q., Peterson, K. R., Fang, X., & Stamatoyannopoulos, G. (2002). Locus control regions. *Blood*, *100*(9), 3077–86. doi:10.1182/blood-2002-04-1104

Liu, E. T., Pott, S., & Huss, M. (2010). Q&A: ChIP-seq technologies and the study of gene regulation. *BMC Biology*, *8*(1), 56. doi:10.1186/1741-7007-8-56

Mandal A. (2015). What is Gene Expression? Retrieved February 13, 2015, from http://www.news-medical.net/health/What-is-Gene-Expression.aspx

Maston, G. a, Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics*, *7*, 29–59. doi:10.1146/annurev.genom.7.080505.115623

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., … Wasserman, W. W. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, *42*(Database issue), D142–7. doi:10.1093/nar/gkt997

Mertes, F., ElSharawy, A., Sauer, S., van Helvoort, J. M. L. M., van der Zaag, P. J., Franke, A., … Brookes, A. J. (2011). Targeted enrichment of genomic DNA regions for next-generation sequencing. *Briefings in Functional Genomics*, *10*(6), 374–386. doi:10.1093/bfgp/elr033

Min, K., Son, H., Lim, J. Y., Choi, G. J., Kim, J.-C., Harris, S. D., & Lee, Y.-W. (2014). Transcription factor RFX1 is crucial for maintenance of genome integrity in Fusarium graminearum. *Eukaryotic Cell*, *13*(3), 427–36. doi:10.1128/EC.00293-13

Missense Prediction Tool Catalogue | NGRL Manchester. (n.d.). Retrieved April 4, 2015, from http://www.ngrl.org.uk/Manchester/page/missense-prediction-tool-catalogue

Mourad Elloumi, A. Y. Z. (2011). Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications. Retrieved February 13, 2015, from http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470505192.html

MutationAssessor.org /// functional impact of protein mutations. (n.d.). Retrieved April 5, 2015, from http://mutationassessor.org/

MutationTaster - documentation. (n.d.). Retrieved February 13, 2015, from http://www.mutationtaster.org/info/documentation.html

Nandi, S., & Ioshikhes, I. (2012). Optimizing the GATA-3 position weight matrix to improve the identification of novel binding sites. *BMC Genomics*, *13*(1), 416. doi:10.1186/1471-2164-13-416

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812–4. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=168916&tool=pmcentrez&rendertype=abstract

Norrgard K, S. J. (2008). Using SNPs Data to Examine Human Phenotypic Differences. Retrieved February 13, 2015, from http://www.nature.com/scitable/topicpage/using-snp-data-to-examine-human-phenotypic-706

NumPy — Numpy. (n.d.). Retrieved February 13, 2015, from http://www.numpy.org/

Pan, Y., Tsai, C., Ma, B., & Nussinov, R. (2009). c o m m e n ta ry How do transcription factors select specific binding sites in the genome ?, *16*(11), 1118–1121.

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews. Genetics*, *10*(10), 669–80. doi:10.1038/nrg2641

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, *20*(1), 110–21. doi:10.1101/gr.097857.109

Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., & Mann, R. S. (2010). Origins of specificity in protein-DNA recognition. *Annual Review of Biochemistry*, *79*, 233–69. doi:10.1146/annurev-biochem-060408-091030

Ruklisa, D., Ware, J. S., Walsh, R., Balding, D. J., & Cook, S. A. (2015). Bayesian models for syndrome- and gene-specific probabilities of novel variant pathogenicity. *Genome Medicine*, *7*(1), 5. doi:10.1186/s13073-014-0120-4

Sample to Insight - QIAGEN. (n.d.). Retrieved February 13, 2015, from http://www.qiagen.com/au/

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., & Hume, D. A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews. Genetics*, *8*(6), 424–36. doi:10.1038/nrg2026

Schwarz, J. M., Rödelsperger, C., Schuelke, M., & Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods*, *7*(8), 575–6. doi:10.1038/nmeth0810-575

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*, *40*(Web Server issue), W452–7. doi:10.1093/nar/gks539

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R., & Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences*, *39*(9), 381–399. doi:10.1016/j.tibs.2014.07.002

Stergachis, A. B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., … Stamatoyannopoulos, J. A. (2013). Exonic transcription factor binding directs codon choice and affects protein evolution. *Science (New York, N.Y.)*, *342*(6164), 1367–72. doi:10.1126/science.1243490

Stormo, G. D. (2013). Modeling the specificity of protein-DNA interactions. *Quantitative Biology*, *1*(2), 115–130. doi:10.1007/s40484-013-0012-4

Strachan T, R. A. (1999). Human Molecular Genetics. Retrieved February 13, 2015, from http://www.ncbi.nlm.nih.gov/pubmed/21089233

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., Kondrashov, A. S., & Bork, P. (2001). Prediction of deleterious human alleles. *Human Molecular Genetics*, *10*(6), 591–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11230178

Taylor, R. W., & Turnbull, D. M. (2005). Mitochondrial DNA mutations in human disease. *Nature Reviews. Genetics*, *6*(5), 389–402. doi:10.1038/nrg1606

Teer, J. K., & Mullikin, J. C. (2010). Exome sequencing: the sweet spot before whole genomes. *Human Molecular Genetics*, *19*(R2), R145–51. doi:10.1093/hmg/ddq333

The UNIX System, UNIX System. (n.d.). Retrieved February 13, 2015, from http://www.unix.org/

Thomas, A. (2013). Thrive in Genetics. Retrieved April 2, 2015, from https://global.oup.com/ushe/product/thrive-in-genetics-9780199694624;jsessionid=033083206D421E7ECE6BF45F452889FF?cc=au&lang=en&

Transcription Factor | Broad Institute of MIT and Harvard. (n.d.). Retrieved February 13, 2015, from https://www.broadinstitute.org/education/glossary/transcription-factor

Transcriptomics | Modeling Immunity. (n.d.). Retrieved February 13, 2015, from http://miepdev.vbi.vt.edu/bioinformatics/transcriptomics/

Wasserman, W. W., & Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews. Genetics*, *5*(4), 276–87. doi:10.1038/nrg1315

Welcome to Python.org. (n.d.). Retrieved February 13, 2015, from https://www.python.org/

Whole Exome Sequencing | Cost-effective analysis of protein coding regions. (n.d.). Retrieved February 24, 2015, from http://www.illumina.com/applications/sequencing/dna_sequencing/exome-sequencing.html

Wilbanks, E. G., & Facciotti, M. T. (2010). Evaluation of algorithm performance in ChIP-seq peak detection. *PloS One*, *5*(7), e11471. doi:10.1371/journal.pone.0011471

Wu, J., & Jiang, R. (2013). Prediction of deleterious nonsynonymous single-nucleotide polymorphism for human diseases. *TheScientificWorldJournal*, *2013*, 675851. doi:10.1155/2013/675851

Yang, F., Zhang, Y., Ressler, S. J., Ittmann, M. M., Ayala, G. E., Dang, T. D., … Rowley, D. R. (2013). FGFR1 is essential for prostate cancer progression and metastasis. *Cancer Research*, *73*(12), 3716–24. doi:10.1158/0008-5472.CAN-12-3274

Yoon, J.-K., Ahn, J., Kim, H. S., Han, S. M., Jang, H., Lee, M. G., … Bang, D. (2015). microDuMIP: target-enrichment technique for microarray-based duplex molecular inversion probes. *Nucleic Acids Research*, *43*(5), e28. doi:10.1093/nar/gku1188

Zhao, Y., Ruan, S., Pandey, M., & Stormo, G. D. (2012). Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, *191*(3), 781–90. doi:10.1534/genetics.112.138685